From Contestation to Camaraderie: Structural Similarity Dampens Toxic Discourse in Polarized Social Groups

Matthew Yeaton
HEC Paris

Sarayu Anshuman
UC Berkeley

Sameer B. Srivastava
UC Berkeley

July 16, 2024

Abstract

Partisan animosity—negative thoughts, feelings, and behaviors towards an outgroup—has been on the rise around the world and has been linked to such societal threats as the weakening of democratic institutions and political violence. It is especially prevalent in online social interactions and frequently results in group members not only disliking one another but even resorting to *toxic discourse*—language that is uncivil, intolerant, or threatening toward others. We develop a novel theoretical account of toxic discourse's antecedents—one that spotlights individuals' positions in the network structure of the group. We hypothesize that, in polarized online groups, the more two individuals are *structurally similar* to one another, the less likely they will be to use toxic language in their interpersonal communication. We further propose that the relationship between structural similarity and toxicity will be moderated by *group polarization*: The more (less) polarized a group is, the more (less) structural similarity will tend to dampen toxic discourse between individuals. To evaluate these ideas, we draw on a rich dataset, encompassing more than 25 million comments made by over 1.7 million users in six polarized communities on Reddit. We build on recent advances in machine learning and network analysis to derive an omnibus measure of interpersonal structural similarity. We validate our measure and then report results from cross-sectional analyses and two natural experiments that provide support for our theory. We discuss implications for research on partisan animosity, group polarization, the measurement of structural similarity, and the interplay of structure and culture.

Recent years have seen striking increases in partisan animosity—that is, negative thoughts, feelings, and behaviors towards an outgroup—around the world and especially in the United States (Iyengar et al., 2019; Finkel et al., 2020; Gidron et al., 2020; Phillips, 2022). Because it has been linked to such societal threats as the breakdown of cross-partisan relationships, the weakening of democratic institutions, and even the eruption of political violence and insurrection, a growing body of interdisciplinary research has examined the causes of, and potential strategies for curbing, partisan animosity (Hartman et al., 2022; Argyle et al., 2023; Davis and Wilson, 2023; Combs et al., 2023; Karell et al., 2023).

Group interactions are especially likely to produce partisan animosity when they occur on digital media, which tend to sort individuals in ways that foment negative intergroup sentiment (Bail, 2022; Törnberg, 2022). In such settings, the absence of non-verbal cues and the opportunity to conceal or blur identities can result in group members not only disliking one another but even resorting to *toxic discourse*—language that is rude, disrespectful, or unreasonable (e.g., Avalle et al., 2024). Exposure to toxic discourse has, in turn, been linked to anxiety, depression, and even to diminished trust in public institutions (e.g., Mutz and Reeves, 2005; Gervais, 2015; Faris et al., 2020). Yet not all groups whose members disagree with or dislike one another engage in toxic discourse. Moreover, in some groups, toxic discourse is episodic or only occurs in pockets. What accounts for variation in who engages in toxic discourse with whom in a social group?

An extensive literature has examined the antecedents of partisan animosity—although only a handful of studies have explored the causes of toxic discourse per se.[1] Prevailing explanations fall into four broad categories (for recent reviews, see Iyengar et al., 2019; Hartman et al., 2022): (1) individual characteristics—for example, holding inaccurate beliefs about an outgroup (e.g., Enders and Armaly, 2019); (2) interactional features—for example, the burstiness of user engagement in digital media (e.g., Avalle et al., 2024); (3) partisan sorting—the growing tendency for ideological identities to align with various social identities (Mason, 2018; Baldassarri and Park, 2020; Bonikowski et al., 2021); and (4) institutional-level shifts in government (e.g., Theriault and Rohde, 2011) and the design of news and social media (Berry and Sobieraj, 2013; Bail, 2022) that engender and amplify toxic discourse.

While acknowledging that the causes of toxic discourse are multifaceted, we propose that the prior literature has overlooked an important explanatory factor: individuals' positions in the network structure of the group. Specifically, we build on—but also importantly deviate

[1]Partisan animosity is part of a cluster of closely related constructs, including affective polarization, political sectarianism, and moral polarization (Finkel et al., 2020; Hartman et al., 2022). We discuss the distinctions among these constructs and our choice to focus on partisan animosity below.

from—a long line of sociological research on structural equivalence, or the extent to which two individuals have similar relations with all other individuals in a network (Lorrain and White, 1971; Doreian, 1988; Burt, 1987). Prior work on structural equivalence has generally assumed that it breeds competition between actors. This literature would predict that the relationship between structurally similar actors will tend to be contentious, thereby increasing their likelihood of having a toxic exchange. Indeed, support for this expectation comes from a recent study that links structural equivalence to the emergence of bullying and harrassment among schoolmates (Faris et al., 2020). Yet research on structural equivalence has assumed that actors are primarily focused on achieving instrumental objectives—for instance, accumulating information (Burt, 1987), gaining power (Friedkin, 1993) or status (Faris et al., 2020), achieving individual career success (Liu et al., 2016), or obtaining the resources needed to survive as a firm (Podolny et al., 1996)—and that actors are aware of the extent to which they are structurally equivalent with others.

In many social interactions, however, and especially those that take in polarized groups on digital media, people have not only instrumental aims but also expressive ones—for example, the need to affirm their identity or obtain social support. Moreover, given the millions of users who participate on a given platform and the tens of thousands who might belong to a subgroup or engage in a given discussion thread, it is implausible to think that people can infer their structural equivalence with others—at least as the literature has heretofore thought about and measured structural equivalence.

In such settings, we propose that *structural similarity* writ large—rather than structural equivalence as narrowly conceptualized in the prior literature—can carry with it observable signals of shared identity, which can in turn counteract the tendency toward contestation and instead promote camaraderie between individuals (Melamed et al., 2020; Ertug et al., 2022; Harrell and Quinn, 2023). Thus, we hypothesize that, in polarized online groups, the more two individuals are structurally similar to one another—independent of their individual differences, the flow of their discussion, and the macro-structural context of the interaction—the less likely they will be to use toxic language in their interpersonal communication.

Of course, groups also vary in the degree to which they are polarized on substantive issues. Moreover, group polarization is not static and can fluctuate naturally through the entry and exit of members, shifts in topics of discussion, and interventions that lead people to moderate their stances or, perhaps inadvertently, take more extreme views (Bail et al., 2018; Waller and Anderson, 2021; Bail, 2022; Combs et al., 2023). We further propose that the relationship between structural similarity and toxicity will be moderated by the

degree of group polarization given that people are especially attuned to identity signals in more polarized contexts (Guilbeault et al., 2018; Dias and Lelkes, 2022; Rawlings, 2022; Guilbeault et al., 2023). Thus, we also hypothesize that the more (less) polarized the group, the more (less) structural similarity will tend to dampen toxic discourse between individuals.

To evaluate these ideas, we draw on a rich data from the online platform, Reddit. Our data include more than 25 million comments made by over 1.7 million users who are part of six groups (i.e., subreddits) that discuss such polarizing topics as politics, cultural norms, science, and labor movements. The groups as a whole and their members vary considerably in their propensity to engage in toxic discourse. To account for this variation, we draw on recent advances in machine learning and network analysis—specifically, node embedding models (Grover and Leskovec, 2016; Zhang et al., 2018; Zhou et al., 2022)—to develop an omnibus measure of interpersonal structural similarity.

We first demonstrate that our measure assesses a facet of interpersonal alignment that differs from a content-based measure of semantic similarity, which—unlike our measure—has become relatively commonplace in sociological research (Kozlowski et al., 2019; Zhou, 2022; Gouvard et al., 2023; Aceves and Evans, 2024). At the same time, our measure is significantly correlated with various facets of network similarity that can be more readily observed by individuals—for example, similarity in network centrality, the degree to which two users overlap in the subreddits they participate in, their overlapping discussion threads, and their shared ties—and that can more plausibly transmit a shared identity signal.

Next we present results from dyad-level, cross-sectional analyses that support our expectation that structural similarity dampens toxic discourse. To establish a causal link between the two, we report results from an instrumental variables regression that takes advantage of a natural experiment in which some dyads experience shifts in their degree of structural similarity based on "rolling greyouts" that randomly and temporarily remove some other users from the platform. Finally, we report analyses from a second natural experiment that induces a shift in the degree to which one of the subreddits in our sample is polarized. Consistent with our theory, a decline in group polarization undercuts the dampening effect of structural similarity on toxic discourse. We discuss implications for research on partisan animosity, group polarization, the measurement of structural similarity, and the interplay of structure and culture.

THEORY

Group Polarization and Partisan Animosity

Although sometimes conflated in popular discourse and in some prior research, group polarization has two distinct flavors: one based on opinions, attitudes, and preferences on substantive issues—for example, hate speech regulation or gun control—and one based on attitudes, thoughts, and behavior toward different subgroups. The former, often referred to as opinion or issue polarization, is defined as "the extent to which opinions on an issue are opposed in relation to some theoretical maximum" (DiMaggio et al., 1996, p. 693). An extensive literature, spanning disciplinary lines, has examined the trends in, and causes of, opinion polarization, finding mixed evidence of the extent to which it has increased over time (DiMaggio et al., 1996; Mouw and Sobel, 2001; Baldassarri and Bearman, 2007; Baldassarri and Gelman, 2008; Layman et al., 2006; Fiorina and Abrams, 2008; Baldassarri and Goldberg, 2014; DellaPosta, 2020; Baldassarri and Park, 2020).

In contrast, partisan animosity, defined as negative thoughts, feelings, or behavior toward an outgroup, is of the latter variety (Hartman et al., 2022). It is closely related to the more established and widely studied construct of affective polarization, or the extent to which one feels more positively toward one's ingroup relative to one's outgroup (Iyengar et al., 2019; Brensinger and Sotoudeh, 2022). The primary difference is that affective polarization takes into account one's feelings toward one's ingroup, whereas partisan animosity does not. Our choice to focus on partisan animosity rather than affective polarization is partly rooted in the fact that, although recent years have seen increases in affective polarization, this trend has been primarily driven by shifts in outgroup, rather than ingroup, sentiment (Gidron et al., 2020; Phillips, 2022).

Other constructs related to partisan animosity include moral polarization, in which "us versus them" distinctions take on the quality of "good versus evil" (Crimston et al., 2022); and political sectarianism, or "the tendency to adopt a moralized identification with one political group and against another" (Finkel et al., 2020, p. 533). The latter includes not only aversion to an outgroup but also the components of othering and moralization. Ultimately, the distinctions between these constructs are subtle, and they can all give rise to our main outcome of interest, toxic discourse. Our primary reason for anchoring our theory on partisan animosity is because it has the most direct relationship to toxic discourse. That said, the arguments we develop below apply to toxic discourse that can arise through any of these interrelated phenomena.

Partisan Animosity and Online Toxic Discourse

Although individuals vary in their propensity to engage productively with disagreeing others (Minson et al., 2020; Xia et al., 2020; Minson and Chen, 2022), cross-partisan interactions can often become rancorous and descend into toxic discourse—that is, language that is rude, disrespectful, or unreasonable (Xia et al., 2020; Pradel et al., 2024; Avalle et al., 2024). Partisan animosity and the tendency toward toxic discourse are amplified when people interact on digital media, which sort people in ways that create an alignment of conflicts along partisan lines (Törnberg, 2022), allow users to blur or conceal their identity (Brubaker, 2020), and often lack strong norms of civility (Coe et al., 2014).

One concrete example of toxic discourse comes from the empirical setting for the present study—the online platform, Reddit—and is excerpted from our dataset (with the expletive masked): "[T]hese protestors are dirty[,] dirty f***ing people." Of course, this is not an isolated example: A recent Pew survey found that over 40% of Americans reported personally experiencing online harassment, a specific manifestation of toxic discourse (Vogels, 2021). Sustained exposure to such language has negative consequences for the individual and the broader social order. For example, exposure to uncivil political speech can induce feelings of anger and aversion (Gervais, 2015). Abusive comments about authors' posts can undermine their credibility (Searles et al., 2020). And, although exposure to political disagreement does not by itself undermine trust in government, it does have this effect when those disagreements are expressed using uncivil language (Mutz and Reeves, 2005). Finally, toxic discourse that takes the form of online hate speech, which derives from an "online hate ecology," has even been linked to violence and youth suicide in the real world (Johnson et al., 2019). Insofar as one aims to counteract such adverse consequences of toxic discourse, one needs a deeper understanding of the conditions under which it arises.

Research the origins of toxic discourse per se is relatively sparse; however, it fits within the more expansive, interdisciplinary literature on the origins of partisan animosity. One category of explanations relates to individual characteristics. For example, since the early 1980s, perceptions of opinion polarization in the U.S. have begun to diverge from actual levels such that people believe that Republicans and Democrats disagree with each other on substantive issues more than they actually do. This misalignment is consequential given that the more one perceives one's own policy views as being misaligned with those of the opposing political party, the more animus one feels toward that outgroup (Enders and Armaly, 2019). In a similar vein, partisans, especially more extreme ones, can develop inaccurate "meta-perceptions"—that is, how one thinks outgroup members view one's ingroup. Having an

exaggerated negative "meta-perception," for example, assuming that outgroup members exhibit more prejudice toward one's ingroup than they actually do, is predictive of more hostile attitudes toward the outgroup (Moore-Berg et al., 2020).

Insofar as studies have examined the antecedents of toxic discourse rather than partisan animosity in general, they have tended to focus on interactional features, particularly in online settings. As a starting point, individuals who are prone toxic discourse are especially likely to self-select into partisan discussion threads. Furthermore, more engaged users, who comment frequently on others' posts, are more likely to use toxic language (Mamakos and Finkel, 2023). Toxic language use breeds more toxic language: Correlational evidence shows that toxic language use in a focal comment increases with the amount of toxic language used in the parent comment (Xia et al., 2020), while experimental evidence suggests that exposure to toxic language in comments increases the toxicity of subsequent comments (Kim et al., 2021). Finally, in a more comprehensive dataset that encompasses eight online platforms (including Facebook, Reddit, Twitter, and YouTube), toxic discourse is more likely to occur when user exchanges become more frequent, discussions veer into controversial topics, and user engagement peaks (Avalle et al., 2024).

Zooming out from individual characteristics and interaction dynamics, a third class of explanations for rising partisan animosity focuses on partisan sorting. In the U.S., for example, the two main political parties are more polarized in their views and are better at sorting people along ideological lines—including on moral, social, and cultural issues (DiMaggio et al., 1996; Baldassarri and Gelman, 2008; Layman et al., 2006; DellaPosta et al., 2015). Although there is a secular trend toward more progressive views on such topics as gay rights and gender roles, it is a partisan shift: Democrats tend to adopt such views before Republicans do (Baldassarri and Park, 2020). These persistent patterns of partisan sorting, particularly on moral issues, provide fuel to the fire of partisan animosity.

Finally, institutional shifts in government and the news and social media reinforce an overall context that fans partisan animosity's flames. In the U.S., for example, government institutions are becoming increasingly polarized, and political elites are especially prone to rancorous exchange (Theriault and Rohde, 2011; Card et al., 2022; Ballard et al., 2023). Meanwhile, social media platforms are designed in ways that motivate people to distort their understanding of themselves and to present the most extreme versions of their selves when communicating with others online (Bail, 2022).

Although we do not discount any of these explanations for the origins of online toxic dis-

course, they are ill-suited to explaining why toxic exchanges occur among some disagreeing users but not others or how they ebb and flow across different pockets of a platform. To explain such phenomena, we contend that it necessary to examine the micro-structural dynamics of online interactions. Specifically, we focus on the similarity of individuals' positions in the network structure of the group.

Structural Similarity and Contestation

Structural similarity is a core thread in multiple sociological literatures—although it goes by different names in each. For example, research in the population ecology tradition has formulated different conceptions of the niche an organization occupies within a resource space and examined its consequences for organizational vitality and mortality (Hannan and Freeman, 1989; McPherson, 1983; Baum and Singh, 1994; Carroll, 1985). The concept of a niche exemplifies a foundational sociological principle that is also central to our theory: There exists a duality between actors and positions, and the positions actors occupy importantly shape their opportunities and constraints (Simmel, 1902; Breiger, 1974; Breiger and Mohr, 2004). In population ecology, structural similarity manifests in the form of *niche crowding* and is assumed to breed competition between organizations that are vying for valuable resources. For example, in their study of the semiconductor industry, Podolny et al. (1996) define niche overlap based on the patent citation network and firms' common dependence on prior inventions. Niche crowding is then defined as the sum of a firm's niche overlaps with all other firms. The authors theorize and find support for the prediction that niche crowding lowers a firm's survival chances.

Whereas population ecology mostly considers structural similarities *between* organizations, a related literature on workforce demography has examined its consequences for individual attainment *within* organizations. For example, similarity can be conceptualized as being part of the same entering cohort. Individuals who join an organization as part of a larger entering cohort face greater competition for opportunities and therefore experience lower rates of career mobility than those who join in a smaller cohort (Stewman and Konda, 1983). Similarly, Burt (1997) finds that the benefits of bridging structural holes decline with the number of people doing the same work as the focal actor. In more recent work based on email data from two disparate organizations, Liu et al. (2016) define structural similarity as the overlap in email distribution lists between two individuals who report to the same supervisor. Distribution list overlap proxies for the degree of competition they face for supervisor attention and financial rewards. The authors find that individuals with higher levels of list overlap receive lower performance ratings and bonuses (Liu et al., 2016).

The assumption that structural similarity breeds competition can be traced back to an even earlier literature in the social networks tradition on *structural equivalence*. In their original conceptualization, Lorrain and White (1971) defined structurally equivalent actors as those with identical relationships to all other actors in a network. Subsequent work relaxed this definition by partitioning the network into blocks and identifying two people as being structurally equivalent if they are assigned to the same block (Breiger, 1976; Doreian, 1988; Burt, 1987; Friedkin, 1993; Greve et al., 2016). Later work considered other forms of equivalence such as that based on the roles that people occupy (Mizruchi, 1993).

Yet, across these disparate studies, the assumption that similarity begets competition, exemplified by Burt (1987, p. 1291), firmly holds: "The structural equivalence model highlights competition between ego and alter. This includes, in the extreme, the competition of people fighting one another for survival but applies more generally to the competition of people merely using one another to evaluate their relative adequacy....The more similar ego's and alter's relations with other persons are—that is, the more that alter could substitute for ego in ego's role relations...the more intense that ego's feelings of competition with alter are...." Faris et al. (2020) extend this intuition to the context of aggressive behavior among classmates in middle school and high school. Using longitudinal survey data, they find that the structural equivalence between two classmates at one point in time is positively related to the likelihood that one will harass or bully the other at a later time.

From Contestation to Camaraderie

The arguments linking structural equivalence to competition are based primarily on an instrumental logic: People who occupy similar structural positions are competing for such valuable resources as information (Burt, 1987), power (Friedkin, 1993), status (Faris et al., 2020), and managerial attention and rewards (Liu et al., 2016). Moreover, prior research has assumed that: (1) actors operate in a relatively closed network; (2) interactions mostly occur face-to-face or at least offline (i.e., not on the internet) among known others; and (3) actors are aware of their structural similarity with all other actors and, guided by the logic of competition, adjust their interpersonal behavior accordingly.

These assumptions start to break down in the context of digital interactions that take place on such platforms as Twitter (now known as X), Threads, and Reddit. In these settings, network boundaries are porous, with users constantly entering and exiting a given group (e.g., a subreddit) or a discussion thread within the group. Moreover, users can mask or blur their identities (Brubaker, 2020). Finally, given that these platforms include hundreds

of millions of users, with tens of thousands of participants in some discussion threads, it is hard to imagine that users would know their structural equivalence to other users based on the measures of equivalence that have been used in prior studies.

To understand how structural similarity relates to toxic discourse in polarized online groups, we posit that it is necessary to: (1) consider not only users' instrumental objectives but also their expressive aims—for example, affirming their identity, affiliating with a group, and obtaining social support (Lin, 2002); and (2) broaden the conception of similarity beyond mere equivalence to the myriad ways in which two people can be structurally similar (e.g., participating in the same communities). Regarding the former, given how many other users they encounter and their churn, as well as the natural tendency for online fora to become highly polarized (Bail, 2022), we expect users in such settings to be highly attuned to identity signals. Support for this view comes from research on social influence that shows how people stop learning from others and cease to update their views on controversial topics when they receive an even very subtle similarity signal—a numerical score of cognitive similarity on a peripherally related word association task—about the other (Guilbeault et al., 2023).

With respect to the latter, we argue that users draw inferences about shared identity with other users based on a variety of (noisy) signals of structural similarity that are observable to them—for example, the subgroups they belong to, the discussion threads they participate in, and the comments they choose to respond to—even if their strict structural equivalence with others is not. Indeed, recent work shows how people use two different kinds of heuristics to develop cognitive maps of large-scale social networks, which allow them to assess their position relative to even unfamiliar others (Son et al., 2021). The first is a simple similarity heuristic based on the features they have in common with known others (e.g., both participate in a conspiracy-focused thread). The other is a more complex heuristic based on feature mapping: Each node is a feature associated with a person (e.g., participation in a conspiracy-focused thread), and edges are the relationships between features (e.g., the tendency for those who participate in a conspiracy-focused thread to also participate in ones advocating for violent protest). People deploy both of these heuristics simultaneously to learn how to encode large-scale networks and draw inferences about unfamiliar others they encounter.

Building on this insight, we propose that, when two structurally similar people encounter one another they will tend to perceive each other has having shared identity—even if they are relative strangers. This perception will, in turn, activate a homophily response (Kovacs and Kleinbaum, 2020; Melamed et al., 2020; Ertug et al., 2022; Harrell and Quinn, 2023) that will counteract the tendency for structural similarity to yield contestation and will instead

promote greater camaraderie. Insofar as structural similarity can be assessed based on the the myriad ways in which two people are similar, including ones that are observable to users (a point to which we return in the section below, "Node Embeddings as a Window into Structural Similarity"), we therefore anticipate:

Hypothesis 1: In the context of polarized online groups, the more two individuals are structurally similar to one another, the less likely they will be to use toxic language in their interpersonal communication.

The Moderating Role of Group Polarization

Our arguments thus far are subject to the scope condition that a group is already polarized on substantive issues. Yet group polarization is typically not static; rather, for a variety of reasons, it can ebb and flow over time. Exogenous events can, for example, produce temporary shifts in group polarization. This was the case for the Twitter group anchored on the hashtag #RefugeesWelcome following the shock of the November 2015 Paris terrorist attack (Barisione et al., 2019).

Similarly, the arrival of certain topics or terms into a discussion forum—for example, the use of racist or sexist terminology or tropes—can induce polarization shifts (Inara Rodis, 2021). In addition, recent years have seen the emergence of various interventions that are designed to de-polarize online groups (Hartman et al., 2022; Bail, 2022). For example, Balietti et al. (2021) report results from a large-scale experiment in which participants were matched to peers with similar demographic traits and shared interests and then exposed to a controversial topic. The intervention increased support for redistributive policies and decreased polarization. In a similar vein, Combs et al. (2023) develop a mobile chat platform on which they run the following experiment: A treatment group is randomly assigned to have an anonymous cross-partisan conversation on a controversial topic, while a control group writes an essay using the same conversation prompts. Results indicate that the intervention resulted in lower levels of polarization in the former group relative to the latter.

Finally, polarization shifts can arise through compositional changes in a community—for example, the departures of more moderate, longtime Twitter users and the arrival of new extremists following Elon Musk's takeover of the platform (Murthy, 2024). In the present study, we similarly take advantage of a natural experiment (described in greater detail below) that induces extreme members of a subreddit to depart and thus drives down opinion polarization among those who remain.

Regardless of how or why a group experiences a change in polarization, we propose that such a shift will influence the degree to which structural similarity activates the mechanisms of contestation versus camaraderie. Given the established link between polarization and identity salience (Guilbeault et al., 2018), we propose that—all else equal—increases (decreases) in group polarization will make people more (less) attuned to the identity signals that can be inferred from observable forms of structural similarity. In other words, the more polarized a group, the more the structural similarity between two individuals will function as a marker of their solidarity. Thus, we expect:

Hypothesis 2: Group polarization will moderate the effects of structural similarity on toxic discourse: The more (less) polarized the group, the more (less) structural similarity will tend to dampen toxic discourse between individuals.

NODE EMBEDDINGS AS A WINDOW INTO STRUCTURAL SIMILARITY

The challenge of measuring structural similarity is longstanding and thorny. Prior work in this vein has employed one of two main approaches. The first involves selecting one or more setting-specific measures that relate to a theoretically relevant facet of structural similarity, while the second entails measuring—particularly through the use of blockmodeling techniques—different notions of structural *equivalence* in networks. Each of these approaches has significant limitations.

Exemplars of the former include various measures of network neighbor overlap—for example, niche crowding between firms based on their overlapping patent citations (Podolny et al., 1996) or competition between workers reporting to the same supervisor based on their overlapping memberships on email distribution lists (Liu et al., 2016). Such measures reveal certain facets of similarity but suffer from two main limitations. First, they are difficult to generalize across settings given that the social or functional roles that matter for consequential outcomes, and thus the relevant dimensions of neighbor overlap, vary by context. Second, such measures naturally emphasize the first-order network to the exclusion of higher-order network ties.

The second approach to measuring structural similarity relies on blockmodeling techniques that were first introduced to sociological research in the early 1970s (Lorrain and White, 1971; Breiger et al., 1975; White and Reitz, 1983; Burt, 1990). Both deterministic blockmodels and stochastic blockmodels (Wang and Wong, 1987; Holland et al., 1983; Anderson et al., 1992; Nowicki and Snijders, 2001) approach the question of structural similarity through

the notion of equivalence classes. These models seek to partition networks according to nodes' patterns of connections. For example, the classic notion of structural equivalence creates partitions that group together nodes that share the exact same set of ties (Lorrain and White, 1971). Regular equivalence creates partitions that group nodes that are equally related to equivalent others (White and Reitz, 1983).

While an emphasis on equivalence relations grounds the blockmodeling approach in strong algebraic foundations that are derived from first principles, the focus on equivalence rather than similarity also introduces problems. Although there are some settings in which strict structural equivalence is necessary to explain a social phenomenon, such cases are the exception rather than the rule. In most cases, modest deviations from equivalence can still be expected to activate the theoretical mechanisms (e.g., competition) as equivalence per se. Moreover, equivalence calculations involve the quantization of a latent, continuous similarity space, which can potentially result in the loss of useful information.

The distinction between equivalence and broader conceptions of similarity comes into sharper focus when we consider the example of a weighted, directed graph rather than an unweighted, undirected graph. Consider the toy example of the small graph, G, in Figure 1. In Panel (a), with an unweighted, undirected graph, nodes i and j are structurally equivalent (in the sense of Burt (1987)). Now suppose that G is a weighted, directed graph. If we introduce a small perturbation in the weights (e.g., either i or j sends more messages to the other), they will no longer be strictly equivalent. Yet we might still expect them to perceive one other as similar actors.

INSERT FIGURE 1 ABOUT HERE

Such problems are magnified in the case of large networks. Structurally equivalent nodes are infrequently observed in real-world networks of any reasonable size even without accounting for the complications of weighted and directed network ties. These problems are partially mitigated by other types of equivalence relations that have been developed in social networks research. For example, role equivalence (Burt, 1990) and automorphic equivalence (Borgatti and Everett, 1989) represent significantly weaker equivalence relations. Yet problems remain even with these alternative measures. Consider again the toy example in Figure 1. In Panel (b), with the addition of k, i and j are neither structurally equivalent, automorphically equivalent, nor role equivalent. However, they nonetheless share many important structural properties that one may want to account for. As noted above, the limitations of equivalence are clear even in this toy example and only get amplified when we consider the scale and

complexity of large, real-world networks.

Indeed, the rarity of true equivalence in observed networks was one of the main factors that drove the development of stochastic, rather than deterministic, blockmodels. Stochastic blockmodeling (SBM) relaxes the strict equivalence restrictions of deterministic blockmodeling by extending the basic framework to assign blocks probabilistically by assuming an underlying latent block structure (Wang and Wong, 1987; Nowicki and Snijders, 2001). This approach maintains many of the desirable conceptual elements of deterministic blockmodeling but weakens structural equivalence into a probabilistic form of similarity. SBM is well-suited to assessing similarities even in the presence of small perturbations to edge weights as described. Yet, SBM is nonetheless based on notions of equivalence that are assumed to be reflected in the underlying latent block structure that probabilistically generated the observed network. In other words, SBM is robust to small perturbations in edge weights but less capable of detecting more nuanced forms of structural similarity.

The disadvantages of equivalence-based approaches become even more apparent when we consider the fact that they focus only on the first-order connections around an individual rather than the broader network structure in which individuals are embedded. Many notions of structure (especially in large and complex networks) would benefit from information that is only observable in higher-order network representations. Returning to our example in Panel (b) of Figure 1, we can see that i and j are more similar in the global structure than they might seem if we only consider their first-order ties (e.g., the fact that j has a tie to k, whereas i does not).

This problem is well-understood in the study of network centrality. For example, degree centrality and eigenvector centrality are related but place differential emphasis on local versus global network properties. Degree centrality strictly considers the first-order network, while eigenvector centrality balances local and global network properties. For some questions, we may prefer degree centrality, however, it is useful to have eigenvector centrality in our arsenal for when global network structure is important to our research question. Better still might be an omnibus measure that reflects the many ways in which two actors might be central in a network.

In thinking about the move from setting-specific and equivalence-based similarity measures to an omnibus one, it is useful to consider the analogous shift that occurred in language models from dictionary-based methods to bag-of-words techniques to semantic similarity measures based on word embeddings. In certain applications, for example, identifying specific instances

of a compound like magnesium sulfate in a scientific abstract, a dictionary-based approach might be most appropriate. Yet such an approach would fail to detect that two scientific abstracts are focused on the same topic (i.e., sulfates) if one mentioned magnesium sulfate, whereas another used the term gypsum (which also happens to be a sulfate). In this case, a topic modeling approach (e.g., Latent Dirichlet Allocation), which infers similarity between documents based on all shared words (without regard to the order in which they appear), might prove more effective (Blei et al., 2003; DiMaggio et al., 2013). Yet topic modeling would be ill-suited to discovering that gypsum is also referred to as hydrated calcium sulfate (in scientific texts) and drywall (in texts related to building and construction). Trained on large corpora from both the scientific and construction domains, embedding models would be better equipped to uncover the semantic similarity between these three terms (Li et al., 2016; Dieng et al., 2020; Grootendorst, 2022; Aceves and Evans, 2024). This is because embedding models take into account the words that appear in the local context (i.e., before and after) around a focal word and thereby learn how to represent words in a high-dimensional space. Words that are proximate in this space are semantically similar to each other.

Given their focus on particular forms of overlapping relations, setting-specific structural similarity measures and deterministic blockmodeling are analogous to dictionary-based approaches to measuring linguistic similarity. Meanwhile, stochastic blockmodeling is a probabilistic model that seeks to uncover latent discrete equivalence categories and shares both conceptual and methodological similarities with topic modeling via Latent Dirichlet Allocation. Both are probabilistic generative models that use latent variables to uncover hidden structures within the data. They employ similar inference techniques for parameter estimation and provide a probabilistic interpretation of the observed data. To draw the final parallel, our preferred approach to assessing structural similarity—node embeddings—is rooted in the same underlying logic as word embeddings.

Node embedding models extend the logic of word embedding models to the realm of social networks (Grover and Leskovec, 2016; Zhang et al., 2018; Zhou et al., 2022). They improve upon both setting-specific and blockmodeling-based approaches in that they take into consideration the broader network structure in which individuals are embedded rather than focusing primarily on the first-order connections around an individual. They also consider a broader array of forms structural similarity can take rather than privileging one measure—e.g., centrality or shared ties—over others. Yet, as we demonstrate below, at high levels of similarity, they effectively proxy for a range of network similarity measures, including ones that are readily observed by individuals. In sum, node embedding models yield an omnibus measure of structural similarity that is more robust to perturbations in relationship char-

acteristics, and more attuned to subtler forms of similarity, than either similarity measures based on specific network attributes or structural equivalence.

Figure 2 illustrates how we operationalize structural similarity based on node embeddings (Panel I) and how these embeddings reveal, using our toy example, the structural similarity between three illustrative nodes (Panel II). Panel I depicts the process our node embedding model uses for all actors in our network. First, it constructs potential sequences that capture meaningful information about nodes' structural positions in the network. Panel I.a shows how these sequences are constructed. We start by selecting node i. Next, we select an alter for i. In this example, we have proceeded from i to j. From here, there are four options we could choose for the path's next step. We choose the next step proportional to the return hyperparameter p (which influences whether we return to i) and the in-out hyperparameter q (which influences whether we move farther away from the prior node i). Higher q will generate context sequences that sample more heavily the local context, while lower q will sample more heavily the distant context. After many such steps, we have a created a single structural context sequence $(i, j, x_2, \ldots)$.

INSERT FIGURE 2 ABOUT HERE

Figure 2, Panel I.b illustrates how we use these context sequences to create the node embeddings. First, we create many of the sequences described in Panel I.a for each of the nodes. Then, we train a skip-gram model based on context sequences that uses shallow neural networks to predict a given node based on its context. For example, in this case we are trying to predict node i based on the structural context around it. This creates a high-dimensional node embedding space, in which each node in the original network is represented by a vector in that space. This process is akin to a word embedding model, which similarly learns how to represent a word in high-dimensional space by learning about its context.

In Figure 2, Panel II, we see how this node embedding space is used to reveal the structural similarity between nodes. Panel II.a shows how three particular nodes from the network are mapped onto the embedding space. Nodes that are close together in the embedding space occupy similar structural positions in the original network. For example, i and x_2 are automorphically equivalent in the original network and are close together in the node embedding space. Therefore, they have high structural similarity to one another. On the other hand, i and x_3 occupy less similar network positions and are consequently less close to each other. Thus, they are less structural similar to one another compared to i and x_2. Finally, Figure 2, Panel II.b shows how we measure structural similarity based on the cosine

similarity between two vectors of interest.

METHOD

Empirical Setting and Data

The empirical setting for our study is the Reddit platform, which is one of the most popular online social network platforms in the world. As of 2024, Reddit hosts approximately 73 million active daily users and 268 million active weekly users[2], making it the fourth most visited website in the United States and the eighth most visited website globally[3].

Reddit is a text-forward platform that facilitates organized user interactions on spaces called subreddits—online communities in which users interact with each other via posts, which initiate discussion of a topic, and comments, which are various users' reactions and opinions about a given post or prior comment. The resulting chain of comments is known as a thread. Many subreddits involve intense exchanges of diverse views, with disagreements sometimes giving rise to toxic exchange.[4]

An important characteristic of the Reddit platform is that each user can view other users' profiles, comment histories, and their interaction patterns. Figure 3 illustrates what information users can observe about other users. Through such information about other users, a focal user can draw some (noisy) inferences about others' identity and some facets of their position in the overall network.

As users interact with each other via comments and posts, they leave a digital trace that can be mapped onto a tree-like network of communications between individuals. Our basic unit of analysis is a comment. Each comment is made by a focal user—the author—in response to a post or comment made by another user—the receiver of that comment. We construct a weighted, directed communication network from these patterns of interactions. Users in this network are represented by nodes. Ties between users in this network are proportional to the number of comments made by a given author to a given receiver, so that a user

[2]Reddit user statistics provided by Backlinko, a search engine optimization firm.

[3]Statistics provided by Semrush, a marketing firm.

[4]An excerpt from user SolariaHues's response to user Wiz_johnny's post, *Can anyone explain what is Reddit and how it works?*, in the 'NewToReddit' subreddit, nicely explains how the platform works: *Here's my orientation guide: Reddit is a collection of communities (subreddits) you can join and participate in, which each have their own rules and culture. It can help to learn about those things for each community before jumping in by checking for rules and lurking for a bit to see what the community is like. Each community is similar to a message board in a way. People make posts, which start a thread and others comment below and start sub-threads.*

who comments frequently to another user has a larger weighted tie to that user. We use this weighted, directed communication network to construct our time-varying, dyadic-level network similarity measures, including our measure of structural similarity.

INSERT FIGURE 3 ABOUT HERE

Given that our theory applies to the context of polarized groups, we selected six subreddits that are among the twenty largest and most active on the platform and that are known to include sharp disagreements among users. We pooled all comments from these six subreddits over a six-month period from January 2022 to June 2022. Table 2 provides descriptive statistics on these six subreddits, which span such diverse topics as politics, science, news, work, and popular culture.

While each of these communities feature sharp disagreement and polarized opinions, not all of them reflect the most traditional notion of polarization along a yardstick of American political polarization from liberal to conservative. While this particular yardstick reflects an important notion of polarization, it is far from the only one. Even within the realm of American political polarization, the liberal-to-conservative spectrum is a truncated representation of the range of American political opinions. For example, we analyze one of the largest online social movement communities oriented around the labor movement and workers' rights. While this community is large, diverse, and polarized, it is more accurate to describe its spectrum as ranging from "anarcho communist" to "liberal" rather than liberal to conservative. Beyond the explicitly political discourse, the communities we analyze are polarized on a wide range of issues including vaccination, climate change, and appropriate social norms.

Of these communtities, the subreddit, 'aita,' in which users discuss and seek feedback on personal arguments they have had, has the largest number of comments (10,190,349), while the subreddit, 'science,', in which users discuss scientific research, has the least (661,382). Across the six subreddits, the number of posts ranges from 11,991 to 181,012, while the number of comments per post ranges from 28 to 168.

INSERT TABLE 2 ABOUT HERE

Variables

Dependent Variable

We use a widely used, pre-trained language model, Detoxify (Hanu and Unitary team, 2020), to measure toxic discourse in the communication between pairs of Reddit users. Detoxify emerged from a coding challenge posted publicly by Jigsaw, a subsidiary of Google that aims to make the internet a safer space. Before this challenge, Jigsaw had launched a project, Conversation AI, to identify negative online content using the Jigsaw Corpus (a database of millions of comments from various online sources annotated by human coders). The Conversation AI initiative resulted in the development of Perspective API, one of the first tools to identify toxicity and hatred in online communication. When supplied with an input text, Perspective API provides a score for toxicity between 0 and 1 (with 1 being the most toxic). This tool was useful but often yielded erroneous classifications and could only provide a score on a single dimension of toxicity. This proved to be problematic given that different platforms have different filtering requirements for offensive content—for example, some might allow for profanities so long as they not hate-filled and directed to specific individuals or groups.

To encourage more innovation, Jigsaw invited developers from across the world to participate in the "Toxic Comment Classification Challenge" on Kaggle.[5] Based on the Jigsaw Corpus shared as part of this challenge, Unitary AI developed a tool, referred to as Detoxify, that improves upon Perspective API by making fewer classification errors and by providing scores for multiple dimensions of toxicity.

——————— INSERT TABLE 4 ABOUT HERE ———————

Toxicity. Given an input text (in our case, a comment), Detoxify provides a score between 0 and 1 (with 1 being the most toxic) for different dimensions of toxicity: *Severe Toxicity*, *Identity Attack*, *Insult*, *Profanity*, *Threat*, and an overarching measure, *Toxicity*. Table 4 provides definitions of each of these dimension. For our main analyses, we focus on the overarching measure as our dependent variable. As reported in the Appendix, our main results are, however, largely consistent when we consider each dimension of toxicity separately.

For each comment, we compute a standardized measure of overall *Toxicity*. We report examples of the most and least toxic comments from our dataset in Table 1. The measure

[5]Kaggle is Google's online data science platform that encourages developers and machine learning engineers to write and share code. More information about this challenge can be found here: https://github.com/praj2408/Jigsaw-Toxic-Comment-Classification

appears to have face validity. For instance, the comment, "I hope all these racist f***s enjoy have no jobs in 20 years after self driving trucks replace their dumb f***ing asses." has a toxicity score of 0.99932, whereas the comment, "This is a good idea; I think I'll look into this. Thank you" has a toxicity score of 0.00051.

INSERT TABLE 1 ABOUT HERE

Independent Variables

Structural Similarity. We start by constructing a weighted, directed communication network of users based on their interactions on Reddit. Nodes in the network represent users, and edges represent the (weighted, directed) communication between users. The directed edge a_{ij} is given in Equation 1 by

$$a_{ij} = \frac{\text{\# of responses from } i \text{ to } j}{\text{total responses to } j} \tag{1}$$

We construct weekly communication networks for each week in our sample period.

For each network, we train network node embeddings that operationalize the idea that nodes occupying similar structural positions in the network should be close together in an embedding space by applying the node2vec model (Grover and Leskovec, 2016). At a high level, the creation of the measure follows these steps: (a) transform the network representation into sequences of nodes that capture meaningful information about the nodes' positions in network structure; (b) use a skip-gram model (shallow neural network) to train an embedding representation of the sequences by predicting a node given its context; (c) calculate the structural similarity between any two given nodes by taking the cosine distance between them in embedding space.

Specifically, the node2vec approach constructs node context sequences to represent nodes' position in network structure. We illustrate an example of this process in Figure 2. In Figure 2, Panel I.a, we begin by choosing a starting node, in this case user i. In this example, we have already proceeded from the node representing user i to the node representing user j. We continue to construct the node context sequences with a random walk that choose its next step according to the return hyperparameter p and the the in-out hyperparameter q. Given that the sequence is positioned at node j, there are four potential nodes we could move to next: i, x_1, x_2, or x_3. The transition probability to go from node j to any other node z in the random walk are proportional to $\pi_{jz} = \alpha(i,z) \cdot a_{iz}$ where the scaling factor α is given in Equation 2 by

$$\alpha(i,z) = \begin{cases} \frac{1}{p} & \text{if } d(i,z) = 0 \\ 1 & \text{if } d(i,z) = 1 \\ \frac{1}{q} & \text{if } d(i,z) > 1 \end{cases} \tag{2}$$

where $d(\cdot, \cdot)$ is the geodesic distance of any two nodes in the network.

In the case of Figure 2, Panel I.a, the walk returns from j to i depending on the return hyperparameter p and proportional to $\alpha = 1/p$. The walk stays a fixed distance away from i by heading to x_1 proportional to $\alpha = 1$. Finally, the walk heads farther away (distance > 1 from i) depending on the in-out hyperparameter q proportional to $\alpha = 1/q$. Higher q will generate random walks that sample more heavily the local context, while lower q will sample more heavily the distant context. The random walk procedure produces a context sequence: e.g., $(i, j, x_2, \ldots)$.[6]

This sequence is analogous to a sentence in a word embedding approach, except instead of capturing information about semantic relationships as we do for word embeddings, we are instead capturing information about structural relationships between nodes. Figure 2, Panel I.b. illustrates that we produce many such random walk context sequences for each node. Using these sequences, we use a skip-gram model to map these sequences into a node embedding space. Each node $i, x_2, x_3, \ldots$ is represented as a vector in this space. Panel II shows how nodes that are close together in the space occupy similar structural roles in the original network. We then calculate the cosine similarity of nodes in the node embedding space to produce the time-varying, dyad-level measure of structural similarity.

Control Variable and Additional Variables for Validation Checks

Because our main models include a battery of fixed effects (e.g., time, subreddit, and dyad), we include only one time-varying control variable, semantic similarity. This variable captures the extent to which the meanings that are expressed in two users' comments are aligned versus misaligned. In contrast, our main measure of interest, structural similarity, is agnostic to the content of users' comments and is instead derived based on patterns of who comments on whose posts.

Semantic similarity is a useful control because the degree of linguistic alignment between two users might influence their propensity to use toxic language with one another—although the direction of this relationship is conceptually unclear. For example, users who communicate using similar linguistic styles might be attracted to one another and thus less likely to behave in toxic ways because of this perceived similarity. Alternatively, they might perceive each other as rivals who are competing for attention or status, thereby increasing the chances that one decides to resort to toxicity. Beyond semantic similarity, we describe below some additional variables we computed to help validate our measure of structural similarity.

Semantic Similarity. To construct this measure, we employ a pre-trained Sentence-BERT model and map each comment's author to a document embedding space. We chose the widely employed

[6]node2vec experiences performance bottlenecks for large networks. For implementation purposes, we utilize fastnode2vec (Abraham, 2020) that is built on node2vec and is known to perform well on large networks due to its efficient usage of memory.

Sentence-BERT model, 'all-mpnet-base-v2,' as it has been shown to generate high-quality embeddings and performs well on semantic similarity tasks (Devlin et al., 2018; Reimers and Gurevych, 2019).

For each comment, we measure the *Semantic Similarity* between the author and the receiver. To do so, we compute the cosine similarity between the embeddings of the author, w_{author}, and the receiver, $w_{receiver}$ (Equation 3). The more similar the linguistic styles of the sender and receiver, the higher values this measure takes. We take the mean of this cosine similarity measure across all comments between all pairs of users in each week. Thus, *Semantic Similarity* is a time-varying (weekly), dyad-level measure.

$$Semantic\ Similarity_{author,receiver} \quad = \quad cossim(w_{author}, w_{receiver}) \tag{3}$$

To assess how our omnibus measure of structural similarity relates to established network measures, as well as platform-specific similarity measures, we computed the following measures:[7] (a) eigenvector centrality (the importance or 'centrality' of a user relative to all other important users in the network); (b) indegree centrality (for a directed network, the fraction of users that direct interactions toward the focal user); (c) outdegree centrality (for a directed network, the fraction of users that are recipients of comments from the focal user).

To construct dyad-level, centrality similarity measures for pairs of users, we compute the normalized absolute difference of each of the three measures described above. We then subtract this value from 1 to obtain a similarity measure for each user pair. CM in Equation 4 refers to the centrality measure (eigenvector, indegree, or outdegree) of a user. If the difference in the CM of two users is low, it indicates greater similarity in the centrality of these users.

$$Centrality\ Based\ Similarity \quad = \quad 1 - \frac{|CM_{author} - CM_{receiver}|}{Range\ of\ all\ CM\ Values} \tag{4}$$

Thus, we obtain three different centrality similarity measures: *Eigenvector Similarity*, *Indegree Similarity*, and *Outdegree Similarity*.

Next, we generate a clustering-based measure of similarity, which is based on the extent to which a given user's connections are mutually interconnected. Clustering similarity is then defined based on the difference in the clustering coefficients of two users. Equation 5 presents the formula used to generate this measure, with CC in Equation 5 referring to the clustering coefficients of the author and receiver.

$$Clustering\ Similarity \quad = \quad 1 - \frac{|CC_{author} - CC_{receiver}|}{Range\ of\ all\ CC\ values} \tag{5}$$

[7] We generate some of the presented measures using the NetworkX python library.

We also define three additional network measures that capture structural similarity on dimensions that are relatively easy to observe on the Reddit platform. These measures reflect different forms of overlap in users' patterns of interactions.

Subreddit Overlap. This measure captures the common subreddits that a given pair of users participates in. For instance, if user A participates in subreddits {S1, S2, S3, S4} and user B participates in subreddits {S3, S4, S5, S6}, then the measure takes a value of $2/6 = 0.333$, as there are two common subreddits between them {S3, S4}.

Thread Overlap. This measure reflects the common threads that two users participate in. For example, if user A participates in threads {T1, T2, T3, T4} and user B participates in the threads {T4, T5, T6}, then the measure takes a value of $1/6 = 0.167$, as there is one common thread between them {T4}.

Shared Ties. This measure represents the common ties between two users. For instance, if user A has interactions with users {S, T, U, V, W} and user B interacts with users {U, V, W, X, Y, Z}, then the measure takes a value of $3/8 = 0.375$, as they share three ties with three users {U, V, W}.

Estimation: Main Analyses

Our main analyses involve regressions of *Toxicity* on *Structural Similarity* with controls and various fixed effects. Specifically, we estimate two OLS models.

$$Toxicity_{i,t} = \alpha D_{i,t} + \beta_1 Structural\ Similarity_{i,t} + \epsilon_{i,t} \quad (6)$$

$$Toxicity_{i,t} = \alpha D_{i,t} + \beta_1 Structural\ Similarity_{i,t} + \beta_2 Semantic\ Similarity_{i,t} + \epsilon_{i,t} \quad (7)$$

where i refers to a comment, t refers to time (week), and $D_{i,t}$ refers to a vector of fixed effects. We estimate the two OLS models under three different specifications. In the first, we do not include any fixed effects. In the second, we include author, receiver, week, and subreddit fixed effects. Finally, in the third, we substitute author and receiver fixed effects with dyad fixed effects. In all cases, to account for the non-independence of observations, we follow prior literature in using two-way clustered standard errors (i.e., clustered by sender and receiver) (Cameron et al., 2011; Kleinbaum et al., 2013; Liu and Srivastava, 2015).

Natural Experiment: Rolling Greyouts

Concerns over endogeneity are ubiquitous in the study of networks. In particular, one might be concerned about the role of self-selection into conversations and networks on the relationship between structural similarity and toxicity. To clarify how structural similarity relates toxicity, we leverage a natural experiment: Rolling back-end outages on Reddit exogenously manipulated the ability of users to access and comment between 02/05/2022-02/15/2022. This outage shifted the

communication network by preventing a more-or-less random set of comments from being posting on the platform. This rolling greyout shock should change structural similarity through its impact on communication networks, but should have no impact on polarization. We instrument structural similarity via this outage to hone in on the part of the variation in structural similarity that is not driven by endogenous user choice.

What we term "rolling greyouts" are more precisely periods of highly elevated error rates on Reddit's back-end servers between the period of 02/05/2022-02/15/2022. Reddit uses the Atlassian suite of software support tools as part of its server stack. When Reddit suffers from problems with their servers, the Atlassian Statuspage tool sends notifications to both internal developers and to a public API (application programming interface). This allows internal developers to know that something has gone wrong and allows external users to know that the problems are on Reddit's side rather than on the end-user's side.

By monitoring the Reddit Statuspage API, we are able to learn about the nature of the back-end problems and their timing. We selected this particular outage because it was fully unplanned (and therefore could not have been anticipated by users) and lasted several days (giving enough time for the rolling outages to affect a large subset of users). Most importantly, it reflected a high and unpredictable increase in error rates during the period (rather than a total outage of the entire site). In effect, the period of elevated back-end errors prevented a more-or-less random set of users from accessing the site and from posting comments on the platform. We use the rolling greyout natural experiment as an instrument for structural similarity. This approach allows us to extract the part of structural similarity which is most affected by the rolling greyouts. Then, we can use the greyout-driven part of the variation in structural similarity in estimating its impact on toxicity.

For the rolling greyout natural experiment to yield a valid instrument for structural similarity, it must (1) have a strong first-stage, such that the rolling greyouts are strongly correlated with structural similarity; and (2) satisfy the exclusion restriction. In our case, the exclusion restriction is that the rolling greyout natural experiment instrument should only affect toxicity through the channel of shifting communication networks and therefore structural similarity. While the strong first-stage can be tested directly, the exclusion restriction can never be tested directly and can only be judged by its plausibility.

Many potential concerns about the exclusion restriction are ameliorated because the rolling greyouts represent an exogenous subtractive shock to communication networks and thus to structural similarity. Yet, at least one important secondary pathway exists that may violate the exclusion restriction: user frustration over the rolling greyouts. If users are annoyed or frustrated about their ability to post and are aware that the problem is on Reddit's side, they may turn that frustration into toxicity directed at other users. This would violate the exclusion restriction. While it is impossible to fully account for this concern, we address it by limiting our analysis to the first two

days of the greyout period. Our reasoning is that at the beginning of this period, users affected by the rolling greyouts will be less likely to be aware of the problems unless they encounter them personally, will have fewer chances to access the website if they do encounter errors personally, and will be less fatigued (and therefore less frustrated) by the errors compared to how they might experience them after a longer period.

We estimate this model using the following two-stage least squares (2SLS) approach:

$$StructuralSimilarity_{i,t} = \alpha D_{i,t} + \beta_{S1} outage_{i,t} + \eta_{i,t} \tag{8}$$

$$Toxicity_{i,t} = \alpha D_{i,t} + \beta_{IV} \left(StructuralSimilarity_{i,t} \mid outage_{i,t} \right) + \varepsilon_{i,t} \tag{9}$$

where i refers to a comment, t refers to time (week), and $D_{i,t}$ refers to a vector of fixed effects that includes author, receiver, and subreddit. We cluster standard errors at the (author, receiver) level.

Natural Experiment: Shift in Group Polarization

To test Hypothesis 2, which postulates that the relationship between structural similarity and toxicity varies as a function of group polarization, we take advantage of a second natural experiment that results in a steep drop in the extent to which one of the subreddits in our data, antiwork, was polarized on issues related to labor reform. Specifically, an impromptu "bombed" interview by a community leader acted as the release of a pressure valve that encouraged members with opposing views to the leader to leave the community en masse and form a separate subreddit. We look at the relationship of structural similarity on toxic discourse among the stayers in the community (i.e., those who never joined the new subreddit). As shown in Figure 4, we leverage this natural experiment via a difference-in-differences analysis, using two non-labor organizing communities as our control.

——————— INSERT FIGURE 4 ABOUT HERE ———————

Antiwork was one of the fastest growing subreddits throughout 2021 and the first quarter of 2022, including being the fastest growing for several months during that period. This growth led to Antiwork becoming one of the five largest subreddits by comment activity volume by the end of 2021 with over 1.7 million members. The scope of the community made it one of the largest single online hubs for the labor social movement.

However, the rapid growth rate was not without conflict. At its root, the community encountered growing pains relating to conflicts about both goals and means between two major factions. In 2015, the community was composed primarily of self-described abolitionist-minded anarcho communists. As of June 2021, the subreddit described itself as "a subreddit for those who want to end work, are curious about ending work, want to get the most out of a work-free life, want more

information on anti-work ideas and want personal help with their own jobs/work-related struggles." The community's FAQ included the following Q&A: "Why Antiwork?"

> "Anti-work has long been a slogan of many anarchists, communists and other radicals. Saying we are anti-job is not quite right because a job is just an activity one is paid for and we are not all against money. 'Anti-labor' makes us sound like we're against any effort at all and we already get that enough as is. (We're not, by the way.)"[8]

This part of the community sought to abolish work as it was commonly construed. Prior to 2019, the subreddit had fewer than 10,000 members. The community transformed into a social movement as it grew to more than 1.7 million members in only two years. The apparent stakes for the community grew alongside its ideological diversity. At this stage, the community began to split into its two major factions. The first one remained committed to the original ideological goals of the community. The second faction that emerged during this period were self-identified liberals, who were substantially more reform-minded and incrementalist than the abolition-minded anarcho communists. The community became increasingly polarized between these two ideological perspectives. Abolitionist-minded users felt that the liberals sought to water down the core ideals of the movement, while liberals felt that the abolitionists were disconnected from concrete, achievable aims. Twitter user _saintdrew discussed their perspective on the evolution of the community on October 25, 2021, saying that

> "I watched in real-time as r/antiwork went from 'abolish wage slavery and transform the world' to 'my boss sucks! I need a nicer boss' and now anarchists are getting yelled at by liberals for pointing out the sub's original purpose lol. It's wild. I'm not against the wave of people finally being emboldened to reject the bs at their workplaces but the attempts to whitewash the meaning of anti-work is rather frustrating. It's ok to not be antiwork yet. Everyone's radicalization process is different. But trying to water down radical ideas to fit liberal sensitivities is not it man."

The increasing polarization and conflict came to a head in January 2022. The increasing prominence of antiwork as an online hub for the labor online social movement led to a rash of high-profile interview requests. Yet the community had no formal leaders. Instead, the moderators of the community, whose role is more akin to a referee than to a leader, deputized themselves as the spokespeople for the antiwork community and movement. Because the community moderators largely dated back to the early days of the community, they took the opportunity to clarify their vision of the anarcho communist orientation of the movement.

[8] https://web.archive.org/web/20220131104428/https://old.reddit.com/r/antiwork/wiki/index

The inciting incident arrived on January 25, 2022. Without community knowledge or approval, the moderator Doreen Ford gave a disastrous interview with Jesse Watters on Fox News. In an interview ostensibly about the growing Antiwork movement, Ford primarily opined on the virtue of laziness, her time spent as a part-time dog walker, and her ambitions to one day teach philosophy. This caused a schism in the antiwork community reacted: the creation that very same day of a new community called workreform, which was intended to better fit the aims of the liberal part of the community. Antiwork hemorrhaged approximately 500,000 users in the following days, most of them leaving to the newly-formed workreform community. Those that remained after the exodus were broadly supporters of the original abolitionist goals of the community.

In leveraging the interview empirically, we examine the change in the impact of structural similarity on toxic discourse among the stayers in the community (i.e., those who never joined the new workreform subreddit). Figure 4 shows a conceptual illustration of our empirical design, which is a natural experiment design operationalized via a difference-in-differences analysis. We use two non-labor organizing communities as our control.

By limiting our analysis to the stayers, we are more easily able to estimate an "apples-to-apples" comparison of the effect of the shift in polarization. In particular, we want to make sure that we not only account for comments by stayers in our comparison but want to include only stayers comments responding to *other stayers*. If community-level polarization had no effect, we have no reason to expect that the relationship between structural similarity and toxicity should change among these stayer-stayer dyads. In contrast, we predict that the reduction in community-level polarization will amplify the contestation mechanism of structural similarity relative to the camaraderie / homophily mechanism because people will become less attuned to identity signals. Thus, we anticipate that a decrease in polarization will counteract the negative relationship between structural similarity and toxicity.

We estimate the following model:

$$\begin{aligned} Toxicity_{i,t} = \alpha D_{i,t} + \beta_1 StructuralSimilarity_{i,t} + \beta_2 AntiworkStayers_{i,t} + \beta_3 PostInterview_{i,t} \\ + \beta_4 StructuralSimilarity_{i,t} \times AntiworkStayers_{i,t} \\ + \beta_5 StructuralSimilarity_{i,t} \times PostInterview_{i,t} \\ + \beta_6 AntiworkStayers_{i,t} \times PostInterview_{i,t} \\ + \beta_7 StructuralSimilarity_{i,t} \times AntiworkStayers_{i,t} \times PostInterview_{i,t} + \varepsilon_{i,t} \quad (10) \end{aligned}$$

where i refers to a comment, t refers to time (week), and $D_{i,t}$ refers to a vector of fixed effects that includes author, receiver, and semantic similarity. We cluster standard errors at the (author, receiver) level.

RESULTS

Descriptive Statistics and Correlations

Table 3 reports descriptive statistics for the main sample, and Table 6 provides correlations among the key variables of interest. Table 2 reports descriptive statistics and the self-descriptions ("About" column) for the communities (subreddits) that we use in our analyses.

Validating the Structural Similarity Measure

We conduct several validation checks of our structural similarity measure. First, Table 5 illustrates the correlation between our measure and other, more common network-similarity measures, as well as setting-specific structural similarity measures (e.g., thread overlap). Our measure is positively correlated with all of these measures except clustering similarity, with which it is negatively correlated. On balance, this pattern is consistent with the idea that our measure can be thought of as an omnibus one that integrates various forms of similarity.

Next, we build on the idea presented in the section above, "Node Embeddings as a Window into Structural Similarity," that many other structural proxy measures reflect a particular type of extreme quantizing of our measure of structural similarity. This quantizing is due to: (1) overweighting the first-order network (relative to our measure); and (2) making very strict decisions about similarity cutoffs (leaning toward strict equivalence). The challenge in large networks is that equivalence between nodes becomes increasingly rare, and therefore measures based on equivlency or first-order networks become too conservative and increasingly uninformative. We illustrate how this problem manifests in practice in Figure 5.

——————— INSERT FIGURE 5 ABOUT HERE ———————

This figure plots the correlation between quantizations of our structural similarity measure at different cutoffs (on the x-axis) with readily observed structural similarity measures (on the y-axis). As we take increasingly conservative quantized transformations of our original continuous measure, our structural similarity measure converges on the various observable, context-specific similarity measures. Taken together, we interpret these validation checks as evidence that our measure can indeed be considered as a stand-in not only for traditional measures of structural equivalence but also for various manifestations of network similarity.

Main Analyses

Table 7 reports the results of our main analyses: regression models of *structural similarity* on *toxicity* in the presence of controls and various fixed effects. In the context of a polarized setting, Hypothesis 1 argues that the more structurally similar two users are, the less they will resort to

having toxic exchanges with one another. In Model (1), which does not include fixed effects, *toxicity* is negatively related to *structural similarity* (β = -0.0756, p < 0.001). In Model (2), we follow a similar specification but add *semantic similarity* as a control variable. Doing so has no appreciable impact on the coefficient for *structural similarity*. In this specification, *semantic similarity* also has a negative and significant coefficient.

Next, we include author, receiver, week, and subreddit fixed effects in Model (3) and Model (4). In both models, the negative relationship between *structural similarity* and *toxicity* remains; however, the magnitude of the effect attenuates considerably. Interestingly, the inclusion of these fixed effects also results in a reversal of the sign of *semantic similarity*, which is now positive and significant.

In addition to the models that include our main variable of interest and the control, we investigate the variation in the coefficient of *structural similarity* in explaining *toxicity*. We define three indicator variables to capture three categories of *structural similarity*: values between the twenty-fifth percentile value and the median, values between between the median and the seventy-fifth percentile, and values greater than the seventy-fifth percentile. We run Model (5) with *semantic similarity* as a control and include all the previously mentioned fixed effects. The coefficient of all three indicator variables is significant and negative; however, the negative relationship between *structural similarity* and *toxicity* is amplified as the structural similarity between dyads increases. This is consistent with our previous analysis suggesting that the correlation between our measure and observable forms of similarity approaches approaches one as we move to the right tail of *structural similarity*.

Lastly, in Model (6) and Model (7), we consider dyad (author-receiver pair) fixed effects instead of author and receiver fixed effects and observe a similar relationship between *structural similarity* and *toxicity*.

Because our results are presented in terms of the relationship between standardized structural similarity and standardized toxicity, they require some translation to make sense of the effect sizes. First, we observe in Table 6 that a standard deviation of raw toxicity (the unit of our DV) is rather large in our setting, $\sim$ 0.3 on a scale from 0 to 1. We can contextualize these results as representing many somewhat more toxic comments, or as representing much more toxic comments. For example, every 2-3 SDs of standardized toxicity represents 0.6-0.9 units of raw toxicity score, which represents one additional comment that is very toxic.

To ground our interpretation, consider these comments that are very close to the median level of toxicity in our sample: "We have 3 puppies already. I am looking for a place to rent and as soon as I find it I am out of here" or "What? Manchin is in WV, not KY. And his voters know exactly what they voted for, and overwhelmingly support him (70% approval rating in the state). He's the only democrat holding statewide office there. I don't like what he's been doing, but it's hard to

argue he isn't liked in the state. And he's a hell of a lot better than any R." Both examples of median toxicity comments are neither particularly positive or particularly negative.

Comments that are are one standard deviation of raw toxicity above these median examples, and therefore are somewhat more toxic than these baseline median examples include "K, stay mad :) I'll keep spending my money that belongs to me however I want. We know that's what you're really upset about" and "This place sounds like a Neoconservative echo chamber." Such comments reflect higher levels of toxicity than the baseline examples, but are nonetheless mild given the empirical context.

Comments that are two standard deviations of raw toxicity above these median examples include "FOX Noise..Tucker,"Clucker",Carlson...is promoting Russia in this...WTF!???" and "Demonstrated proof that sometimes your elected officials who you think give a shit about you couldn't actually care less and are just in it for themselves. She could have been VP once upon a time." Finally, comments that are three standard deviations of raw toxicity above these median examples include, "It's quite obvious you don't care how it affects your daughter. I hope she moves in with her dad on her 18th birthday and never speaks to you again. You are a horrible mother" and "And yet Fled Cruz would still lick Tweetle Dumb's boots regardless." These examples of comments that are two or three standard deviations above the median are undoubtedly toxic.

Based on this interpretation, and if we consider Model (4) as our core baseline model, we see that a user would make 1.5-2 fewer very toxic comments per 100 comments to an alter with structural similarity one SD above the mean compared to an alter with structural similarity one SD below the mean. At the rate of commenting activity of an average user in our sample, this means that a dyad with structural similarity one SD above the mean would make three to four fewer highly toxic comments to each other per month compared to a dyad with structural similarity one SD below the mean.

Similarly, we would expect that a user would make 4.6-6 fewer somewhat toxic comments per 100 comments to an alter with structural similarity one SD above the mean compared to an alter with structural similarity one SD below the mean, and a dyad with structural similarity one SD above the mean would make nine to twelve fewer highly toxic comments to each other per month compared to a dyad with structural similarity one SD below the mean.

Over two million comments are made on Reddit per day. Even in our sample of communities, there are about 4.1 million comments made per month by hundreds of thousands of users. Thus, returning to Model (4), it is plausible to estimate that Reddit users would make 10,000-20,000 fewer very toxic comments per million comments to alters with structural similarity one SD above the mean compared to alters with structural similarity one SD below the mean.

INSERT TABLE 7 ABOUT HERE

Natural Experiment: Rolling Greyouts

Table 9 presents results of an instrumental variable (IV) analysis leveraging our rolling greyouts natural experiment. Model (1) shows results from a baseline OLS specification for this sample period, and confirms that the basic cross-sectional relationship between *structural similarity* and *toxicity* is negative and significant within this sample period, consistent with the findings of our main cross-sectional analysis. Model (3) shows the results of the first-stage estimation. The rolling greyouts have a significant effect on *structural similarity*, and the first-stage F-statistic is 225.28 which is much larger than the Stock et al. (2002) rule of thumb of $F > 10$. Taken together, these results suggest that the rolling greyout instrument is adequately powered, and is a strong first-stage instrument for *structural similarity*. Model (2) presents the second-stage estimation results, which show a strong, significant negative relationship between *structural similarity* and *toxicity*. This suggests that the part of *structural similarity* that is affected by the rolling greyout shocks maintains the negative relationship with *toxicity* that we have observed in our cross-sectional results.

INSERT TABLE 9 ABOUT HERE

One concern about the validity of the rolling greyout natural experimental design is whether users would be able to perceive changes to their networks as a result of the greyout shocks. We assess the impact of the rolling greyouts on one difficult to observe and two more easily observed measures of network similarity: *structural equivalence* (Model 4); *neighbor overlap* (Model 5); and *thread overlap* (Model 6). We find that the greyout shock has a significant impact on each, suggesting that users should be able to perceive some changes in their networks as a result of the shock.

While it is typical for IV estimates to be two to three times larger than OLS estimates (Lal et al., 2024), we observe IV estimates closer to twenty times as large as our OLS estimates. The large size differences between the OLS and IV estimates may arise from three potential sources: (1) attenuation bias present in the OLS estimate; (2) differences in accounting for sample-selection in the OLS estimate versus the IV estimate; or (3) the possibility that our instrument somehow violates the exclusion restriction (an assumption that cannot be directly tested).

Attenuation bias in OLS estimates leads to coefficient estimates which are biased toward zero relative to the "true" parameter. IV estimates typically exhibit less attenuation bias than OLS estimates because of the two-stage design. This difference is magnified in the case of heterogeneous treatment effects, the variance of which shrinks coefficient estimates (even if those heterogeneous effects have the same sign). Given that we have already presented evidence that points to heterogeneous treatment effects by structural similarity quartile, we suspect that there is some attenuation bias in our OLS estimate.

We expect that the functional difference between the effective sample in our OLS estimation and the effective sample used in our IV estimation is relevant in our context because we expect there to be significant self-selection into conversations and significant heterogeneity across users and across dyads. These are likely to be major sources of variance in the OLS estimates, and this could further account for differences between the two models given that we do not account for this sample selection in our OLS estimates directly.

Given the nature of our rolling greyout natural experiment instrument, we suspect that we are estimating a treatment effect for active users whom we expect to make more attempts to access the platform and thus successfully post their comments. In other words, we believe that our treatment effect is for the subset of users who are more engaged, more active, and more persistent in commenting than the overall population of Reddit. Consistent with this view, we find that users in affected dyads comment three times as much in the pre-outage period compared to others. This interpretation does not affect the validity of the instrumental variable design, but it does affect how we should interpret its main results.

One reason that the effect size of structural similarity on toxicity might be larger for active users is that they are more likely to be aware of and pay attention to their (observable) structural similarity with other users. For example, relative to the casual user, they are more likely to observe that they and another user are both active in the same subreddits or threads. Although we cannot rule out the third possibility of an exclusion restriction violation, we note that our OLS and IV estimates are both negative and significant, consistent with Hypothesis 1.

Natural Experiment: Shift in Group Polarization

Table 8 reports the results of a difference-in-differences analysis exploiting an exogenous shift in group polarization. This analysis allows us to test Hypothesis 2, which proposes that the relationship between structural similarity and toxicty should be moderated by group polarization. Model (1) presents the results of the baseline specification of structural similarity on toxicity in order to confirm the main, cross-sectional result for this sample period. The coefficient on structural similarity is negative and significant, consistent with our prior results. Model (2) introduces semantic similarity into the specifications, and we see that this has little effect on the coefficient for structural similarity. Model (3) introduces Antiwork Stayers and Post-Interview into the specification. The main effect of Antiwork Stayers is statistically indistinguishable from zero, suggesting that there is not a major difference in toxicity between the Antiwork Stayers and the control group of non-labor related community users. The main effect of Post-Interview is positive and significant.

Model (4) presents the results of the main specification, which includes the interaction terms. First, we note that the main effects are similar between Model (3) and Model (4). The coefficient on Structural Similarity $\times$ Antiwork Stayers $\times$ Post-Interview is positive and significant, which

suggests that the shift in group polarization had a significant impact on the relationship between structural similarity and toxicity amongst the Antiwork Stayers (relative to the control group of non-labor related community users), consistent with Hypothesis 2.

We highlight that the effect of the polarization shock on toxicity appears to only manifest *through its effects on structural similarity*. For example, there is no statistical effect of Antiwork Stayers × Post-Interview, which suggests that Antiwork Stayers do not simply increase their toxicity homogeneously. Instead, they appear to differentially increase their toxicity toward structurally similar alters.

Figure 6 presents the marginal effects of structural similarity on toxicity for the treatment group of Antiwork Stayers and the control group of non-labor community users. First, we observe that there do not appear to be significant pre-trends in the pre-shock period in either the treatment or control groups. Additionally, we note that there appears to be no significant difference between the treatment and control groups in the pre-shock period. Next, we observe that there is a significant increase in the effect of structural similarity on toxicity among the treatment group in the two weeks following the shift in group polarization, before attenuating in the third week. Indeed, the marginal effects are *positive* in the first two weeks, suggesting an attenuation strong enough that it may actually lead to structural similarity promoting contestation *over* camaraderie.

Overall, our findings in this natural experiment represent evidence in favor of Hypothesis 2, suggesting that the more polarized the group, the more structural similarity appears to dampen toxic discourse between individuals. When the level of group polarization decreased, the dampening effect weakened. The effect of structural similarity on toxicity became more positive, consistent with a competition and contestation interpretation.

Supplemental Analyses and Robustness Checks

Hyperparameter Tuning

One of the hyperparameters in the node2vec algorithm is the 'q' parameter. This 'in-out parameter' dictates whether the random walk places greater emphasis on a local view of the network (i.e., $q > 1$) versus a global view (i.e., $q \leq 1$). With higher values of 'q,' the algorithm places greater emphasis on increasing the likelihood that users who share neighbors will tend to be closely located in the embedding space.[9] The resulting embeddings emphasize local connections relative to global ones.

With lower values of 'q,' the algorithm places greater emphasis on ensuring that users with similar overall global connections tend to be located closely in the embedding space.[10] Such an approach

[9]This type of exploration in a network mimics a popular algorithm known as 'Breadth First Search.'
[10]This type of exploration in a network mimics an algorithm known as 'Depth First Search.'

emphasizes a global view of a user's neighborhood.

For our main analyses, we set 'q' = 3 as it emphasizes a local view of a user's connections and better corresponds to traditional conceptions of structural equivalence. As a robustness check, however, we select a two-week period in the month of March 2022, and generate the structural similarity measure using different values of 'q' ranging from 0.5 to 3. As shown in Appendix Table A1, we find that across different values of 'q,' the resulting structural similarity measures are highly correlated, indicating that our findings are not an artifact of the tuning of this hyperparameter.

Alternative Measures of Structural Similarity

Given that our structural similarity measure is highly correlated with network similarity and setting-specific similarity measures, one question that arises is whether one could simply use one of these measures in place of our omnibus measure. To evaluate the extent to which an omnibus measure provides a distinctive similarity signal, we estimated the specification in Table 7, Model (4) but with alternative measures of structural similarity substituted in for node embedding-based one. To conserve computational resources, we did so for a subsample of the data: a two-week period in March 2022. These results are presented in Appendix Table A2. Model (1) presents our structural similarity measure based on node embeddings as a reference. The coefficient for *structural similarity* is still negative and significant somewhat different from that in Table 7, Model (4) because we are using a different sample.

Models (2) through (8) consider the alternative measures of structural similarity. Models (2), (3), and (4) consider centrality-based similarity measures. We find that the coefficients are positive, rather than negative, but not significant. Model (5), which employs the similarity measure based on the clustering coefficient, is negative but not significant. Turning to the setting-specific similarity measures, in Models (6) and (8), the coefficients for subreddit overlap and shared ties are both negative but not significant. Only Model (7), which focuses on thread overlap, shows a negative and significant relationship with toxic discourse. If we were choosing ex ante among these alternative measures of structural similarity, there would be no obvious reason to prefer thread overlap to any of the other measures. Thus, we see this analysis as reinforcing the value of having an omnibus measure of structural similarity that integrates all of these different similarity dimensions.

The Potentially Confounding Role of Bots

Reddit is known to have significant activity by bots, a program that pretends to be a real user by generating specific kinds of posts and comments (Hurtado et al., 2019). Despite strict rules and regulations on subreddits, bots are gradually contributing more and more to Reddit traffic. We assess the robustness of our findings to the exclusion of identified bots.

Bots are often identified by having suspicious patterns of behavior. For example, bots tend to reply to comments within seconds. They also have very low comment histories and usually do not have profile pictures. Their comments might also appear odd or exaggerated.

Although bot detection algorithms exist, bot designers are always finding new ways to elude them. To assess the extent to which bot activity might account for our results, we sought to eliminate from our data the users that are most likely to be bots. Specifically, we obtained a list of the fifty most commonly observed bots on Reddit.[11] These bots are active in various subreddits. For example, included in this list is one of the most common bots on the Reddit platform, 'Automoderator,' which moderates user activity to ensure that users are following the regulations of a given subreddit regulations.

The bot list we used was generated by using a time filter to remove comments made within sixty seconds of a parent post/comment. Since it can be tricky and tedious to identify whether every individual user present in our dataset is a bot, we choose a conservative method as a next step to ensure that we have a bot-free dataset: we remove all those comments made by users having the phrase 'bot' in their username. These two filters result in a 2% reduction in the original dataset.

We present the cross-sectional analysis using this refined dataset in Table A3. The results are largely consistent with those reported in Table 7. Recognizing that our procedure is unlikely to have purged our dataset of all bot activity, we nevertheless conclude that our results do not appear to be an artifact of bot-generated content.

Examining Different Dimensions of Toxicity

Our overall measure of toxic language encompasses multiple language dimensions, as outlined in Table 4. Table A4 shows the OLS models in which we use as the dependent variable each of the six dimensions of *toxicity* in place of the overall measure. Model (1) considers the overarching dimension of toxicity as the main dependent variable (the same model as in Model (4) of Table 7). Models (2) through (6) present the OLS models for the other dimensions of *toxicity* as dependent variables. All the models include author, receiver, week and subreddit fixed effects, and we also include semantic similarity as a control. The pattern of results is fairly consistent across all six dimensions.

DISCUSSION

The goal of this article has been to deepen our understanding of the causes of partisan animosity, which is on the rise around the world and poses a significant threat to democratic institutions and

[11]The source for the list of fifty bots is https://www.reddit.com/r/dataisbeautiful/comments/9mh3pn/oc_the_50_most_active_bots_on_reddit_based_on/.

the social order (e.g., Hartman et al., 2022). One of partisan animosity's most potent manifestations is in the form of toxic discourse, which not only fuels more animosity but also makes it harder to heal past divisions. Whereas prior work has linked toxic discourse to various individual differences, features of the conversations that partisans engage in, and institutional shifts in government and the news and social media (e.g., Bail et al., 2018), we instead trace its origins to actors' positions in network structure. Our theory centers on the notion of structural similarity, which permeates through multiple subfields of sociology (Lorrain and White, 1971; Carroll, 1985; Liu et al., 2016; Faris et al., 2020). Departing from the prior literature, however, we broaden the conceptualization of structural similarity beyond mere equivalence to encompass many different forms of similarity, including ones that are more readily observed by peers.

While much of the prior literature on structural similarity views it through the prism of contestation and competition (e.g., Burt, 1987), we propose that—in the context of polarized online groups—it can also promote camaraderie and signal shared identification by virtue of its more observable forms. We first hypothesize that interpersonal structural similarity would restrain two users from using toxic language with one another. We further theorize that group polarization moderates the relationship between structural similarity and toxic discourse such that the more polarized a group is on substantive issues, the more structural similarity will tamp down interpersonal toxic exchanges. To evaluate these ideas, we develop a measure of structural similarity based on a node embedding model (e.g., Grover and Leskovec, 2016). Using a rich dataset of comments exchanged by over 1.7 million users in six polarized groups on Reddit, we find support for our theory in analyses of both cross-sectional data and two natural experiments.

Contributions

Insights from this study contribute to four distinct literatures. First, given that toxic discourse arises through and subsequently fuels not only partisan animosity but also other closely constructs such as affective polarization (Iyengar et al., 2019), moral polarization (Crimston et al., 2022), and political sectarianism (Finkel et al., 2020), we extend the understanding of the microfoundations of these phenomena. Existing theory about these forms of division is better suited to explaining variation in their behavioral manifestations (e.g., toxic discourse) across platforms, groups, and users; however, it is ill-suited to explaining why such behavior emerges in pockets and why it rises and falls among the same group of users over time. Our results fill in some of the missing pieces: Users within a given group are more or less apt to use toxic language with each other as a function of their (changing) structural similarity, and groups vary over time in their degree of issue polarization.

These insights point to new potential interventions that could help curb toxic discourse in online groups. For example, they can help platform designers develop better predictive models of where toxic discourse is most likely to emerge and among which specific users. This might alert moderators

of an online group to take preventative measures to keep the conversation from going off the rails. Next, when making recommendations about whom a given user should connect to in order to discuss a controversial topic (e.g., Combs et al., 2023), platform designers could suggest other users who have differing points of view but are structurally similar to the focal user. Alternatively, platform designers could simply make the structural similarity between users more salient, for example, by highlighting the number of group affiliations, discussion threads, or ties they have in common. Such an intervention could be timed to coincide with an increase in opinion polarization with the group, when it is most likely to be effective. Our results suggest that any of these interventions could prove effective in reducing toxic exchange.

Second, we contribute to research on the linkages between opinion polarization and partisan animus. Prior work has established this relationship via longitudinal survey data and survey experiments. An example of the former is a study documenting that opinions in the U.S. on such issues as social welfare have become increasingly divided along party lines and that social welfare ideology is correlated with partisan animosity (Webster and Abramowitz, 2017). Such associations are tantalizing but not causal. The latter is exemplified by recent work that experimentally manipulates the ideological extremity of hypothetical pairs of candidates and then assesses affective evaluations of the candidates (Rogowski and Sutherland, 2016). Such designs yield causal estimates but lack external validity. We add to this line of work by: (a) showing via the natural experiment of the sudden splintering of a highly polarized subreddit, #antiwork, into two less polarized communities that group polarization is causally linked to partisan animosity; and (b) demonstrating that the relationship between group polarization and partisan animosity has an important dyad-level moderator in the form of interpersonal structural similarity.

Third, this study informs the wide range of sociological research that aims to measure similarities between actors—whether in the context of immigrant assimilation (Schachter, 2016), the cultural fit of individuals in an organization (Goldberg et al., 2016), board interlocks between firms (Mizruchi, 1989), varieties of nationalism in a country (Bonikowski and DiMaggio, 2016), or the schemas that Republicans and Democrats in the U.S. hold about poverty (Hunzaker and Valentino, 2019) and the very notion of "America" (van Loon et al., 2024). Whereas each of these prior studies introduced new approaches to measuring different facets of similarity between actors, we import into sociological research a network-analytic method that captures the myriad ways in which two actors occupy similar positions in social structure. This method promises to do for sociological network analysis what word embedding models have done for cultural analysis—namely, to harness the tools of machine learning to efficiently assess the semantic similarities between sets of words based on their relative distances. We introduce and demonstrate the applicability of a parallel method that can provide an omnibus measure of structural similarity and that can be readily applied to any domain in which the relationships among actors or elements can be represented as a network.

Finally, our study contributes to research on the interplay between structure and culture. It highlights pathways through which the network structure individuals are located in can shape the culture of groups they belong to. For example, online groups can develop norms of toxicity that can become entrenched over time and are robust to the entry and exit of users (Rajadesingan et al., 2020). Some of the interventions described above (e.g., making the structural similarity between users more salient to each of them) could help to reset corrosive group norms. In addition, the development of an omnibus measure of structural similarity that is distinct from semantic similarity (recall that, in our setting, the two measures are not highly correlated and seem to be picking up on distinct forms of similarity) points to the potential of future research that examines how the two relate to each other and how they jointly shape consequential outcomes.

Limitations and Directions for Future Research

Our study is not without limitations, which also point to directions for future research. First, our data are captured almost immediately after a user posts them; however, some of the most toxic posts might have been subsequently removed by subreddit moderators. This is not necessarily problematic, as the target of a toxic post will still have been exposed to the problematic language even if it is later removed. But it is difficult to tell from our data whether a particular toxic post had adverse effects just on its target or more broadly in the community. We leave to future research the task of separating out the two effects. On a related note, although we assume that toxic discourse causes harm to target users and the user community, we do not directly observe those harms. Future research could combine archival analyses of the kind we conducted with survey data from platform users.

A second data quality issue relates to the presence of bots, which are known to be prevalent on platforms such as Reddit (Hurtado et al., 2019). Although we reported a robustness check in which we removed posts from users that had been clearly identified as bots, we recognize that bot designers and the developers of bot detection algorithms are playing an ongoing cat-and-mouse game: As bot detection algorithms improve, more sophisticated bots are continually deployed. It is unclear how to fully address this concern in this line of research.

Finally, although our dataset encompasses over 1.7 million users and six subreddits, it is unclear the extent to which our findings would generalize to other communities on Reddit and to other platforms. A useful next step would be to replicate these analyses across more platforms and communities and to directly measure their changing levels of polarization. This would provide an even stronger test of our second hypothesis.

Conclusion

Occupying the same position in social structure has heretofore been assumed to promote contestation between actors. Yet, when groups become polarized and people seek to affiliate with similar others, that same similarity in network position can transform contestation into camaraderie. Understanding when and how this shift occurs offers new possibilities for salving the weeping wounds of partisan animosity.

References

Abraham, L. (2020). Competition analysis on the over-the-counter credit default swap market. *arXiv preprint arXiv:2012.01883*.

Aceves, P. and J. A. Evans (2024). Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science. *Organization Science 35*(3), 788–814.

Anderson, C. J., S. Wasserman, and K. Faust (1992). Building stochastic blockmodels. *Social Networks 14*(1-2), 137–161.

Argyle, L. P., C. A. Bail, E. C. Busby, J. R. Gubler, T. Howe, C. Rytting, T. Sorensen, and D. Wingate (2023). Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences 120*(41), e2311627120.

Avalle, M., N. Di Marco, G. Etta, E. Sangiorgio, S. Alipour, A. Bonetti, L. Alvisi, A. Scala, A. Baronchelli, M. Cinelli, et al. (2024). Persistent interaction patterns across social media platforms and over time. *Nature 628*(8008), 582–589.

Bail, C. (2022). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

Bail, C. A., L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences 115*(37), 9216–9221.

Baldassarri, D. and P. Bearman (2007). Dynamics of political polarization. *American Sociological Review 72*(5), 784–811.

Baldassarri, D. and A. Gelman (2008). Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology 114*(2), 408–446.

Baldassarri, D. and A. Goldberg (2014). Neither ideologues nor agnostics: Alternative voters' belief system in an age of partisan politics. *American Journal of Sociology 120*(1), 45–95.

Baldassarri, D. and B. Park (2020). Was there a culture war? partisan polarization and secular trends in us public opinion. *The Journal of Politics 82*(3), 809–827.

Balietti, S., L. Getoor, D. G. Goldstein, and D. J. Watts (2021). Reducing opinion polarization: Effects of exposure to similar people with differing political views. *Proceedings of the National Academy of Sciences 118*(52), e2112552118.

Ballard, A. O., R. DeTamble, S. Dorsey, M. Heseltine, and M. Johnson (2023). Dynamics of polarizing rhetoric in congressional tweets. *Legislative Studies Quarterly 48*(1), 105–144.

Barisione, M., A. Michailidou, and M. Airoldi (2019). Understanding a digital movement of opinion: The case of# refugeeswelcome. *Information, Communication & Society 22*(8), 1145–1164.

Baum, J. A. and J. V. Singh (1994). Organizational niches and the dynamics of organizational mortality. *American Journal of Sociology 100*(2), 346–380.

Berry, J. M. and S. Sobieraj (2013). *The outrage industry: Political opinion media and the new incivility*. Oxford University Press.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022.

Bonikowski, B. and P. DiMaggio (2016). Varieties of american popular nationalism. *American Sociological Review 81*(5), 949–980.

Bonikowski, B., Y. Feinstein, and S. Bock (2021). The partisan sorting of "america": How nationalist cleavages shaped the 2016 us presidential election. *American Journal of Sociology 127*(2), 492–561.

Borgatti, S. P. and M. G. Everett (1989). The class of all regular equivalences: Algebraic structure and computation. *Social Networks 11*(1), 65–88.

Breiger, R. L. (1974). The duality of persons and groups. *Social Rorces 53*(2), 181–190.

Breiger, R. L. (1976). Career attributes and network structure: A blockmodel study of a biomedical research specialty. *American Sociological Review*, 117–135.

Breiger, R. L., S. A. Boorman, and P. Arabie (1975). An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology 12*(3), 328–383.

Breiger, R. L. and J. W. Mohr (2004). Institutional logics from the aggregation of organizational networks: Operational procedures for the analysis of counted data. *Computational & Mathematical Organization Theory 10*, 17–43.

Brensinger, J. and R. Sotoudeh (2022). Party, race, and neutrality: investigating the interdependence of attitudes toward social groups. *American Sociological Review 87*(6), 1049–1093.

Brubaker, R. (2020). Digital hyperconnectivity and the self. *Theory and Society 49*, 771–801.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology 92*(6), 1287–1335.

Burt, R. S. (1990). Detecting role equivalence. *Social Networks 12*(1), 83–97.

Burt, R. S. (1997). The contingent value of social capital. *Administrative Science Quarterly 42*(2), 339–365.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics 29*(2), 238–249.

Card, D., S. Chang, C. Becker, J. Mendelsohn, R. Voigt, L. Boustan, R. Abramitzky, and D. Jurafsky (2022). Computational analysis of 140 years of us political speeches reveals more positive

but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences 119*(31), e2120510119.

Carroll, G. R. (1985). Concentration and specialization: Dynamics of niche width in populations of organizations. *American Journal of Sociology 90*(6), 1262–1283.

Coe, K., K. Kenski, and S. A. Rains (2014). Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication 64*(4), 658–679.

Combs, A., G. Tierney, B. Guay, F. Merhout, C. A. Bail, D. S. Hillygus, and A. Volfovsky (2023). Reducing political polarization in the united states with a mobile chat platform. *Nature Human Behaviour 7*(9), 1454–1461.

Crimston, C. R., H. P. Selvanathan, and J. Jetten (2022). Moral polarization predicts support for authoritarian and progressive strong leaders via the perceived breakdown of society. *Political Psychology 43*(4), 671–691.

Davis, D. W. and D. C. Wilson (2023). Stop the steal: Racial resentment, affective partisanship, and investigating the january 6th insurrection. *The ANNALS of the American Academy of Political and Social Science 708*(1), 83–101.

DellaPosta, D. (2020). Pluralistic collapse: The "oil spill" model of mass opinion polarization. *American Sociological Review 85*(3), 507–536.

DellaPosta, D., Y. Shi, and M. Macy (2015). Why do liberals drink lattes? *American Journal of Sociology 120*(5), 1473–1511.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dias, N. and Y. Lelkes (2022). The nature of affective polarization: Disentangling policy disagreement from partisan identity. *American Journal of Political Science 66*(3), 775–790.

Dieng, A. B., F. J. Ruiz, and D. M. Blei (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics 8*, 439–453.

DiMaggio, P., J. Evans, and B. Bryson (1996). Have american's social attitudes become more polarized? *American Journal of Sociology 102*(3), 690–755.

DiMaggio, P., M. Nag, and D. Blei (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics 41*(6), 570–606.

Doreian, P. (1988). Equivalence in a social network. *Journal of Mathematical Sociology 13*(3), 243–281.

Enders, A. M. and M. T. Armaly (2019). The differential effects of actual and perceived polarization. *Political Behavior 41*, 815–839.

Ertug, G., J. Brennecke, B. Kovács, and T. Zou (2022). What does homophily do? a review of the consequences of homophily. *Academy of Management Annals 16*(1), 38–69.

Faris, R., D. Felmlee, and C. McMillan (2020). With friends like these: Aggression from amity and equivalence. *American Journal of Sociology 126*(3), 673–713.

Finkel, E. J., C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, et al. (2020). Political sectarianism in america. *Science 370*(6516), 533–536.

Fiorina, M. P. and S. J. Abrams (2008). Political polarization in the american public. *Annual Review of Political Science 11*, 563–588.

Friedkin, N. E. (1993). Structural bases of interpersonal influence in groups: A longitudinal case study. *American Sociological Review*, 861–872.

Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics 12*(2), 167–185.

Gidron, N., J. Adams, and W. Horne (2020). *American affective polarization in comparative perspective*. Cambridge University Press.

Goldberg, A., S. B. Srivastava, V. G. Manian, W. Monroe, and C. Potts (2016). Fitting in or standing out? the tradeoffs of structural and cultural embeddedness. *American Sociological Review 81*(6), 1190–1222.

Gouvard, P., A. Goldberg, and S. B. Srivastava (2023). Doing organizational identity: Earnings surprises and the performative atypicality premium. *Administrative Science Quarterly 68*(3), 781–823.

Greve, H. R., J.-Y. Kim, and D. Teh (2016). Ripples of fear: The diffusion of a bank panic. *American Sociological Review 81*(2), 396–420.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Grover, A. and J. Leskovec (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864.

Guilbeault, D., J. Becker, and D. Centola (2018). Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences 115*(39), 9714–9719.

Guilbeault, D., A. van Loon, K. Lix, A. Goldberg, and S. B. Srivastava (2023). Exposure to the views of opposing others with latent cognitive differences results in social influence—but only when those differences remain obscured. *Management Science*.

Hannan, M. T. and J. Freeman (1989). *Organizational ecology*. Harvard University Press.

Hanu, L. and Unitary team (2020). Detoxify. Github. https://github.com/unitaryai/detoxify.

Harrell, A. and J. M. Quinn (2023). Shared identities and the structure of exchange distinctly shape cooperation. *Social Forces 102*(1), 223–241.

Hartman, R., W. Blakey, J. Womick, C. Bail, E. J. Finkel, H. Han, J. Sarrouf, J. Schroeder, P. Sheeran, J. J. Van Bavel, et al. (2022). Interventions to reduce partisan animosity. *Nature Human Behaviour 6*(9), 1194–1205.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

Hunzaker, M. F. and L. Valentino (2019). Mapping cultural schemas: From theory to method. *American Sociological Review 84*(5), 950–981.

Hurtado, S., P. Ray, and R. Marculescu (2019). Bot detection in reddit political discussion. In *Proceedings of the fourth international workshop on social sensing*, pp. 30–35.

Inara Rodis, P. d. C. (2021). Let's (re) tweet about racism and sexism: responses to cyber aggression toward black and asian women. *Information, Communication & Society 24*(14), 2153–2173.

Iyengar, S., Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science 22*, 129–146.

Johnson, N. F., R. Leahy, N. J. Restrepo, N. Velásquez, M. Zheng, P. Manrique, P. Devkota, and S. Wuchty (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature 573*(7773), 261–265.

Karell, D., A. Linke, E. Holland, and E. Hendrickson (2023). "born for a storm": Hard-right social media and civil unrest. *American Sociological Review 88*(2), 322–349.

Kim, J. W., A. Guess, B. Nyhan, and J. Reifler (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication 71*(6), 922–946.

Kleinbaum, A. M., T. E. Stuart, and M. L. Tushman (2013). Discretion within constraint: Homophily and structure in a formal organization. *Organization Science 24*(5), 1316–1336.

Kovacs, B. and A. M. Kleinbaum (2020). Language-style similarity and social networks. *Psychological science 31*(2), 202–213.

Kozlowski, A. C., M. Taddy, and J. A. Evans (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review 84*(5), 905–949.

Lal, A., M. Lockhart, Y. Xu, and Z. Zu (2024). How much should we trust instrumental variable estimates in political science? practical advice based on 67 replicated studies. *Political Analysis*, 1–20.

Layman, G. C., T. M. Carsey, and J. M. Horowitz (2006). Party polarization in american politics: Characteristics, causes, and consequences. *Annual Review of Political Science 9*, 83–110.

Li, C., H. Wang, Z. Zhang, A. Sun, and Z. Ma (2016). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 165–174.

Lin, N. (2002). *Social capital: A theory of social structure and action*, Volume 19. Cambridge university press.

Liu, C. C. and S. B. Srivastava (2015). Pulling closer and moving apart: Interaction, identity, and influence in the us senate, 1973 to 2009. *American Sociological Review 80*(1), 192–217.

Liu, C. C., S. B. Srivastava, and T. E. Stuart (2016). An intraorganizational ecology of individual attainment. *Organization Science 27*(1), 90–105.

Lorrain, F. and H. C. White (1971). Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology 1*(1), 49–80.

Mamakos, M. and E. J. Finkel (2023). The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS Nexus 2*(10), pgad325.

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago Press.

McPherson, M. (1983). An ecology of affiliation. *American Sociological Review*, 519–532.

Melamed, D., M. Sweitzer, B. Simpson, J. Z. Abernathy, A. Harrell, and C. W. Munn (2020). Homophily and segregation in cooperative networks. *American Journal of Sociology 125*(4), 1084–1127.

Minson, J. A. and F. S. Chen (2022). Receptiveness to opposing views: Conceptualization and integrative review. *Personality and Social Psychology Review 26*(2), 93–111.

Minson, J. A., F. S. Chen, and C. H. Tinsley (2020). Why won't you listen to me? measuring receptiveness to opposing views. *Management Science 66*(7), 3069–3094.

Mizruchi, M. S. (1989). Similarity of political behavior among large american corporations. *American Journal of Sociology 95*(2), 401–424.

Mizruchi, M. S. (1993). Cohesion, equivalence, and similarity of behavior: A theoretical and empirical assessment. *Social Networks 15*(3), 275–307.

Moore-Berg, S. L., B. Hameiri, and E. Bruneau (2020). The prime psychological suspects of toxic political polarization. *Current Opinion in Behavioral Sciences 34*, 199–204.

Mouw, T. and M. E. Sobel (2001). Culture wars and opinion polarization: the case of abortion. *American Journal of Sociology 106*(4), 913–943.

Murthy, D. (2024). Sociology of twitter/x: Trends, challenges, and future research directions.

Annual Review of Sociology 50.

Mutz, D. C. and B. Reeves (2005). The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review 99*(1), 1–15.

Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association 96*(455), 1077–1087.

Phillips, J. (2022). Affective polarization: Over time, through the generations, and during the lifespan. *Political Behavior 44*(3), 1483–1508.

Podolny, J. M., T. E. Stuart, and M. T. Hannan (1996). Networks, knowledge, and niches: Competition in the worldwide semiconductor industry, 1984-1991. *American Journal of Sociology 102*(3), 659–689.

Pradel, F., J. Zilinsky, S. Kosmidis, and Y. Theocharis (2024). Toxic speech and limited demand for content moderation on social media. *American Political Science Review*, 1–18.

Rajadesingan, A., P. Resnick, and C. Budak (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 14, pp. 557–568.

Rawlings, C. M. (2022). Becoming an ideologue: Social sorting and the microfoundations of polarization. *Sociological Science 9*, 313–345.

Reimers, N. and I. Gurevych (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rogowski, J. C. and J. L. Sutherland (2016). How ideology fuels affective polarization. *Political Behavior 38*, 485–508.

Schachter, A. (2016). From "different" to "similar" an experimental approach to understanding assimilation. *American Sociological Review 81*(5), 981–1013.

Searles, K., S. Spencer, and A. Duru (2020). Don't read the comments: the effects of abusive comments on perceptions of women authors' credibility. *Information, Communication & Society 23*(7), 947–962.

Simmel, G. (1902). The number of members as determining the sociological form of the group. i. *American Journal of Sociology 8*(1), 1–46.

Son, J.-Y., A. Bhandari, and O. FeldmanHall (2021). Cognitive maps of social features enable flexible inference in social networks. *Proceedings of the National Academy of Sciences 118*(39), e2021699118.

Stewman, S. and S. L. Konda (1983). Careers and organizational labor markets: Demographic models of organizational behavior. *American Journal of Sociology 88*(4), 637–685.

Stock, J. H., J. H. Wright, and M. Yogo (2002). A survey of weak instruments and weak identi-

fication in generalized method of moments. *Journal of Business & Economic Statistics 20*(4), 518–529.

Theriault, S. M. and D. W. Rohde (2011). The gingrich senators and party polarization in the us senate. *The Journal of Politics 73*(4), 1011–1024.

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences 119*(42), e2207159119.

van Loon, A., A. Goldberg, and S. B. Srivastava (2024). Imagined otherness fuels blatant dehumanization of outgroups. *Communications Psychology 2*(1), 39.

Vogels, E. A. (2021). The state of online harassment. *Pew Research Center 13*, 625.

Waller, I. and A. Anderson (2021). Quantifying social organization and political polarization in online platforms. *Nature 600*(7888), 264–268.

Wang, Y. J. and G. Y. Wong (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association 82*(397), 8–19.

Webster, S. W. and A. I. Abramowitz (2017). The ideological foundations of affective polarization in the us electorate. *American Politics Research 45*(4), 621–647.

White, D. R. and K. P. Reitz (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks 5*(2), 193–234.

Xia, Y., H. Zhu, T. Lu, P. Zhang, and N. Gu (2020). Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction 4*(CSCW2), 1–23.

Zhang, D., J. Yin, X. Zhu, and C. Zhang (2018). Network representation learning: A survey. *IEEE transactions on Big Data 6*(1), 3–28.

Zhou, D. (2022). The elements of cultural power: novelty, emotion, status, and cultural capital. *American Sociological Review 87*(5), 750–781.

Zhou, J., L. Liu, W. Wei, and J. Fan (2022). Network representation learning: from preprocessing, feature extraction to node embedding. *ACM Computing Surveys (CSUR) 55*(2), 1–35.

Tables and Figures

Table 1: Examples of High and Low Toxicity Comments. This table presents representative examples of high-toxicity and low-toxicity Reddit comments. The examples shown are all from the top 50 or bottom 50 for predicted toxicity in the sample (the top and bottom 0.0001^{st} percentile). Beyond illustrating the face validity of our toxicity measure, these examples also illustrate that debate and discussion do not intrinsically lead to toxicity. Several of the low toxicity comments shown here reflect constructive and friendly responses to political discussions (including on the topics of vaccination and climate change).

	Raw Toxicity	Comment from the Reddit Dataset
High Toxicity	0.99932	I hope all these racist f***s enjoy having no jobs in 20 years after self driving trucks replace their dumb f***ing asses
	0.99908	#f*** nestle they are child murdering hyper capitalist monsters
	0.99923	Psycho f**ing maniacs with badges taking out their self hate on dogs
	0.99921	F***in pussy leftists won't even arm themselves
	0.99913	F*** joe Biden f*** The Clintons f*** Obama f*** the left the libs and honestly f*** the grimy ass republicans too
	0.99919	F*** nazis and f*** tankies
	0.99916	these protestors are dirty dirty f***ing people
	0.99915	Mayor can suck my socialist c***
	0.99914	These f***ing clowns are desperate to paint themselves as victims
	0.99912	dont f***ing pardon them keep the terrorists in jail were they f***ing belong to be honest throw some corrupted DAs in there too for f*** ing up the process
Low Toxicity	0.00049	I think that last sentence would be fascinating to investigate. Thank you.
	0.00048	Thank you for your kind words, I just wanted to share the wisdom.
	0.00047	Cheers, thank you for the links.
	0.00050	Thank you for the detailed response. I'm not well read on the topic so this was very informational.
	0.00048	Thank you for this perspective, I think I'm going to extend an olive branch and say that I could have done better and we will go from there. Thanks for taking the time to write this up.
	0.00051	This is a good idea; I think I'll look into this. Thank you.
	0.00049	I'm glad these kind of studies are being done. It is important that all aspects of cannabis consumption should be researched and information given to the public who can make an informed choice, if or when they decide to do so.
	0.00049	Very interesting, thank you for sharing your own experience. Would be interesting to investigate such variations granted the vaccine assists, but from an aspect which investigates the reasons why some people react so poorly when others do not experience severe illness. I have seen some studies looking at genetics and also previous exposure to to various illnesses.
	0.00049	I read a different report that predicted we could be as far as 10 feet underwater by 2100 if we continue the way we are. This document you shared has some awesome information. Thank you for sharing.
	0.00049	Study after study shows that diversity improves group productivity and decision-making, so I am happy to see it as an addition to the qualification process.

Table 2: Description of Subreddits

Subreddit	No. of comments	No. of distinct authors	No. of posts	Mean no. of comments per post	Mean Toxicity	About
aita	10,190,349	652,705	181,012	56.297	0.178	A catharsis for the frustrated moral philosopher in all of us, and a place to finally find out if you were wrong in an argument that's been bothering you. Tell us about any non-violent conflict you have experienced; give us both sides of the story, and find out if you're right, or you're the asshole.
news	2,719,733	417,603	16,220	167.678	0.151	The place for news articles about current events in the United States and the rest of the world. Discuss it all here.
politics	4,416,997	480,235	47,985	92.049	0.147	Politics is for news and discussion about U.S. politics.
science	661,382	212,194	11,991	55.156	0.054	This community is a place to share and discuss new scientific research. Read about the latest advances in astronomy, biology, medicine, physics, social science, and more. Find and submit new publications and popular science coverage of current research.
worldnews	7,254,960	744,208	74,948	96.799	0.133	A place for major news from around the world, excluding US-internal news.
antiwork	975,278	129,863	34,964	27.894	0.160	Antiwork: Unemployment for all, not just the rich! A subreddit for those who want to end work, are curious about ending work, want to get the most out of a work-free life, want more information on anti-work ideas and want personal help with their own jobs/work-related struggles.

Table 3: Description of the Sample Used for the Cross-sectional Analysis

No. of comments	25,243,421
No. of distinct authors	1,795,724
Average length of comments	41.987 (words)
Average Toxicity	0.153

Table 4: Dimensions of Toxicity in the Annotated Jigsaw Corpus

1.	Toxicity	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion
	a. Severe Toxicity	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.
	b. Identity Attack	Negative or hateful comments targeting someone because of their identity.
	c. Insult	Insulting, inflammatory, or negative comment towards a person or a group of people.
	d. Profanity	Swear words, curse words, or other obscene or profane language.
	e. Threat	Describes an intention to inflict pain, injury, or violence against an individual or group.

Table 5: Correlations of structural similarity with standard network measures

Network Similarity Measure	Mean	S.D.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
(1) Structural Similarity	0.585	0.218	–							
(2) Indegree Similarity	0.945	0.143	0.206***	–						
(3) Outdegree Similarity	0.987	0.073	0.130***	-0.033***	–					
(4) Eigenvector Similarity	0.982	0.079	0.109***	0.530***	0.018***	–				
(5) Clustering Similarity	0.998	0.016	-0.112***	-0.035***	-0.009***	-0.016***	–			
(6) Subreddit Overlap	0.067	0.076	0.090***	0.098***	0.048***	0.060***	-0.004***	–		
(7) Thread Overlap	0.111	0.181	0.282***	-0.040***	0.066***	-0.001***	-0.047***	0.093***	–	
(8) Shared Ties	0.008	0.045	0.078***	0.052***	0.008***	-0.012***	-0.075***	0.498***	0.254***	–

Note: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$

Table 6: Descriptive Statistics and Correlation Matrix

Variables	Mean	S.D.	**1**	**2**	**3**
(1) Toxicity	0.153	0.290	–		
(2) Structural Similarity	0.628	0.222	-0.076***	–	
(3) Semantic Similarity	0.382	0.177	-0.007***	0.028***	–

Note: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$

Table 7: OLS Regressions of Toxicity on Structural Similarity and Semantic Similarity

Dependent Variable:				Toxicity			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variables							
Structural Similarity	-0.0756***	-0.0755***	-0.0190***	-0.0194***		-0.0127***	-0.0127***
	(0.0082)	(0.0082)	(0.0006)	(0.0005)		(0.0031)	(0.0031)
Semantic Similarity		-0.0045**		0.0326***	0.0326***		0.0118***
		(0.0017)		(0.0006)	(0.0006)		(0.0025)
StructuralSimilarity_greater75					-0.0456***		
					(0.0015)		
StructuralSimilarity_medianto75					-0.0252***		
					(0.0011)		
StructuralSimilarity_25toMedian					-0.0133***		
					(0.0009)		
Fixed-effects							
Author			Yes	Yes	Yes	Yes	Yes
Receiver			Yes	Yes	Yes		
Week			Yes	Yes	Yes		
Subreddit			Yes	Yes	Yes	Yes	Yes
Dyad						Yes	Yes
Fit statistics							
Observations	25,243,421	25,243,421	25,243,421	25,243,421	25,243,421	25,243,421	25,243,421
R^2	0.00572	0.00574	0.23406	0.23480	0.23478	0.90961	0.90962

Two-way standard errors (clustered by sender and receiver) in parentheses.
*Note: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$*

Table 8: Difference-in-Differences Analysis: Exogenous Shift in Group Polarization.

Dependent Variable:	Toxicity			
Model:	(1)	(2)	(3)	(4)
Variables				
Structural Similarity	-0.0177***	-0.0183***	-0.0184***	-0.0173***
	(0.0015)	(0.0015)	(0.0015)	(0.0020)
Semantic Similarity		0.0499***	0.0499***	0.0498***
		(0.0019)	(0.0019)	(0.0019)
Antiwork Stayers			0.0025	0.0034
			(0.0111)	(0.0119)
Post-Interview			0.0134***	0.0152***
			(0.0035)	(0.0038)
Structural Similarity × Antiwork Stayers				0.0039
				(0.0039)
Structural Similarity × Post-Interview				-0.0078**
				(0.0025)
Antiwork Stayers × Post-Interview				-0.0109
				(0.0091)
Structural Similarity × Antiwork Stayers				0.0131*
× Post-Interview				(0.0053)
Fixed-effects				
Author	Yes	Yes	Yes	Yes
Receiver	Yes	Yes	Yes	Yes
Fit statistics				
Observations	2,284,679	2,284,679	2,284,679	2,284,679
R^2	0.39536	0.39657	0.39657	0.39658

Two-way standard errors (clustered by sender and receiver) in parentheses.
*Note: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$*

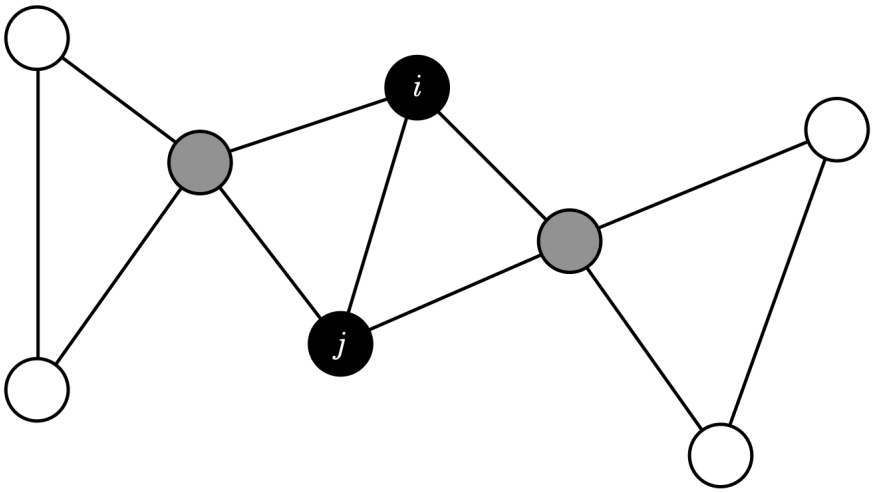
Table 9: Instrumental Variable Analysis of a Rolling Greyout Natural Experiment on the Relationship Between Structural Similarity and Toxicity

Dependent Variables:	Toxicity		Structural Similarity	Structural Equivalence	Neighbor Overlap	Thread Overlap
	OLS	2SLS (2nd Stage)	2SLS (1st Stage)			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Structural Similarity	-0.0185***	-0.5671**				
	(0.0014)	(0.1892)				
Rolling Greyouts			-0.0394***	0.0331***	-0.0165*	-0.0222***
			(0.0089)	(0.0080)	(0.0078)	(0.0055)
Fixed-effects						
Author	Yes	Yes	Yes	Yes	Yes	Yes
Receiver	Yes	Yes	Yes	Yes	Yes	Yes
Subreddit	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	2,704,606	2,704,606	2,704,606	2,704,606	2,704,606	2,704,606
F-test (1st stage)			225.28			
R^2	0.37943	0.31807	0.79611	0.59461	0.49188	0.83731

Two-way standard errors (clustered by sender and receiver) in parentheses.
*Note: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$*

Figure 1: Conceptual Figure Illustrating Various Notions of Structural Similarity Between the Nodes i and j in Graph G.

(a) This figure shows the small, toy graph G in which nodes i and j are embedded. In this initial example, i and j are structurally equivalent (in the sense of Burt (1987)).



(b) This figure shows another version of the toy graph G. The addition of node k has a large impact on the first-order network similarity of i and j, yet we may still be interested in the extent of structural similarity between them.

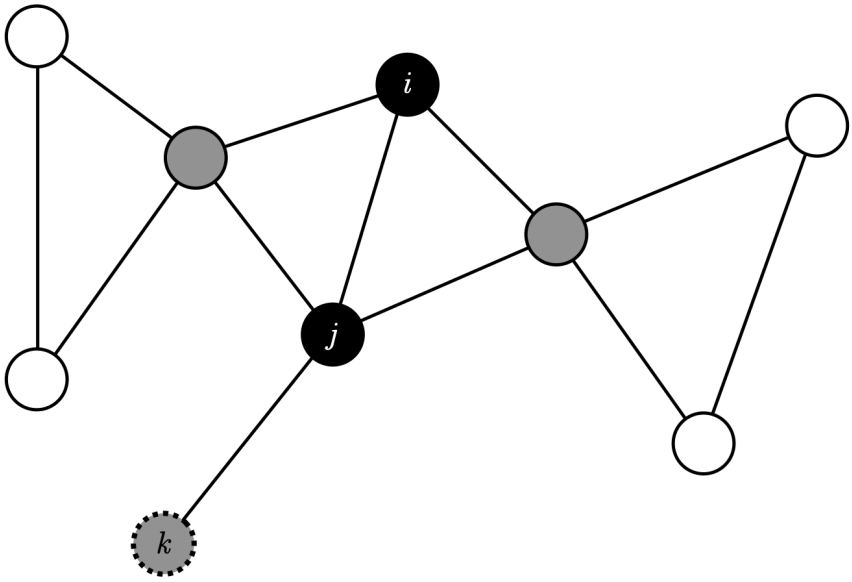


Figure 2: Conceptual Figure of Structural Similarity Measure. Panel I. illustrates the process used to create the node embeddings for all nodes in the network. First, I.a. shows how we create structural context sequences that capture meaningful information about nodes' structural positions in the network. In this example, we have proceeded from i to j. From here, there are four options the path could choose for its next step, and we choose the next step proportional to the hyperparameters p and q. Fixing a given p, higher q generates context sequences that sample more heavily the local context, while lower q will sample more heavily the distant context. Iteratively, this process creates the structural context sequence $(i, j, x_2, \ldots)$. I.b. illustrates how we use these context sequences to create the node embeddings. We create many of the sequences described in I.a. for each of the nodes. Then, we train a skip-gram model to predict a given node based on its context. This creates a high-dimensional node embedding space, where each node in the original network is represented by a vector in that space. In Panel II, we can see how this node embedding space is used to reveal the structural similarity between nodes. II.a. shows how three nodes from the network are mapped into the embedding space. We can see that nodes that are close together in the node embedding space occupy similar structural roles in the original network. Finally, II.b. shows how we can measure structural similarity using cosine similarity.

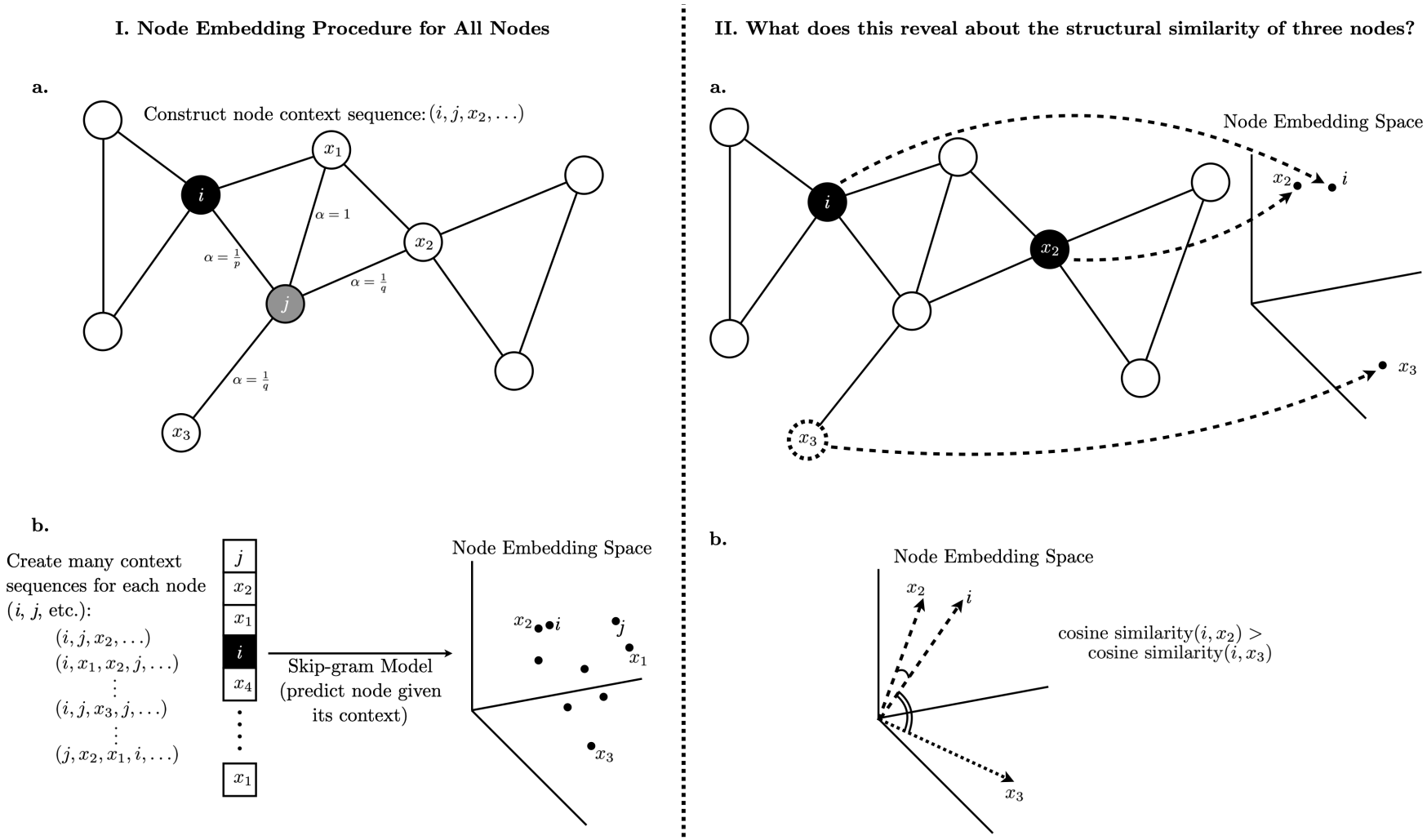


Figure 3: Example reddit user-discovery interface: Users can click on the username of their interlocutor. This brings them to their user page. From here, much can be quickly discerned about the activity of that other user, including past comment history (including past comment locations) sorted by various activity metrics. In this example, the focal user selects the user page of nov4chip. From here, the focal user can browse their past recent activity, for example scrolling back to see comments in soccer and Inter Milan.

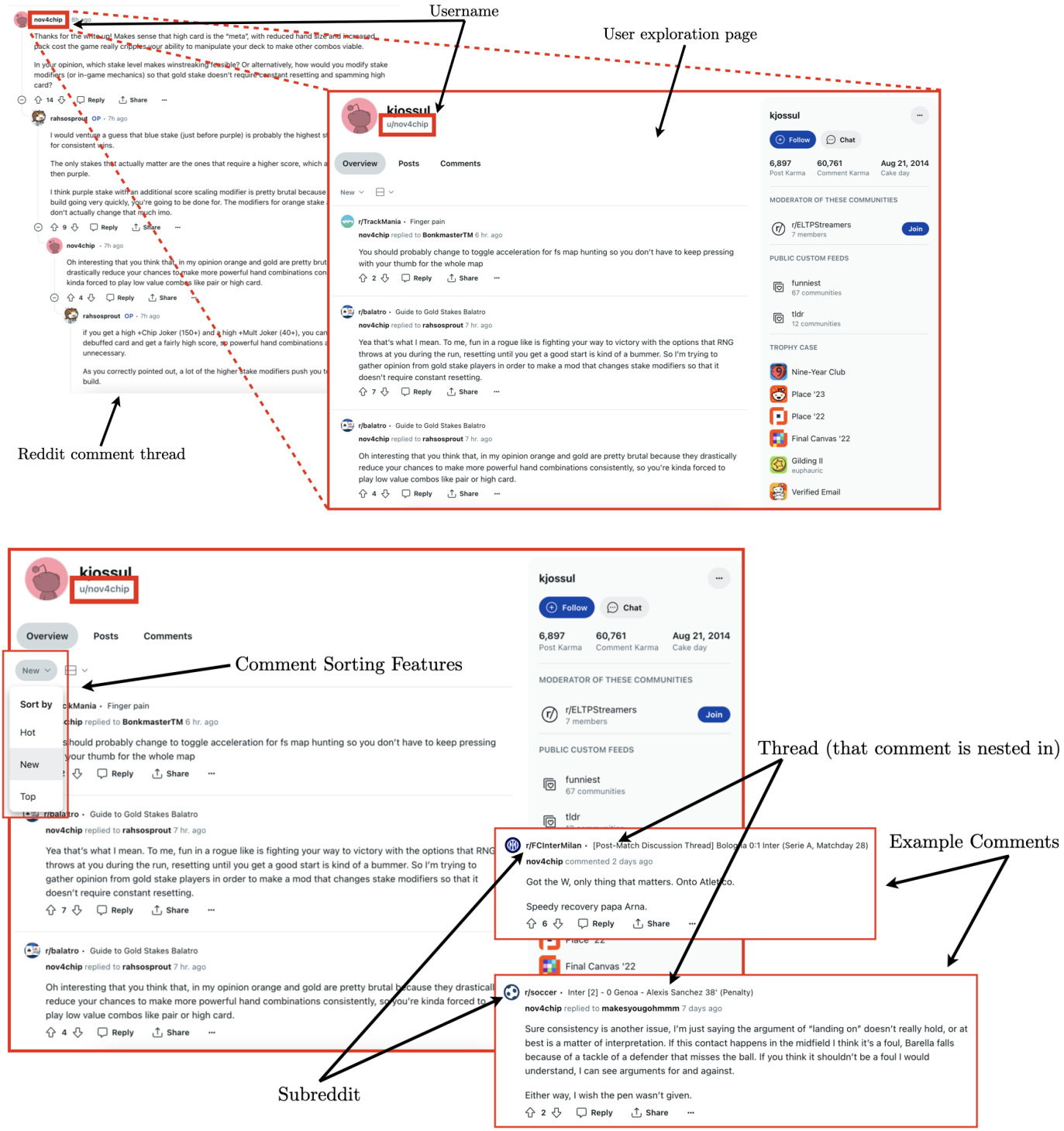


Figure 4: Conceptual Representation of Natural Experiment Leveraging a Shift in Polarization

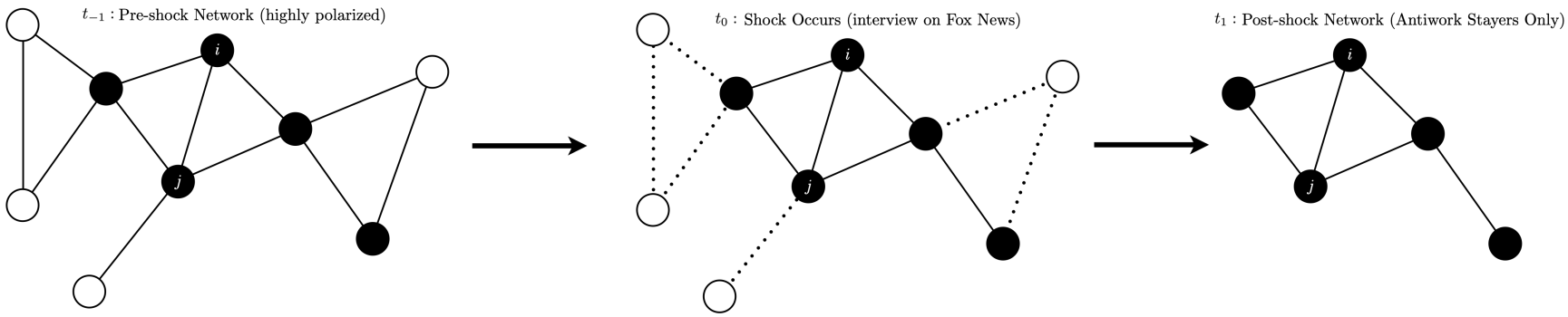


Figure 5: Other Common-sense Structural Proxy Measures Capture Narrow Notion of Structural Similarity that is Rarely Reflected in Large Networks: Correlation between Quantized Structural Similarity and Other Structural Proxy Measures

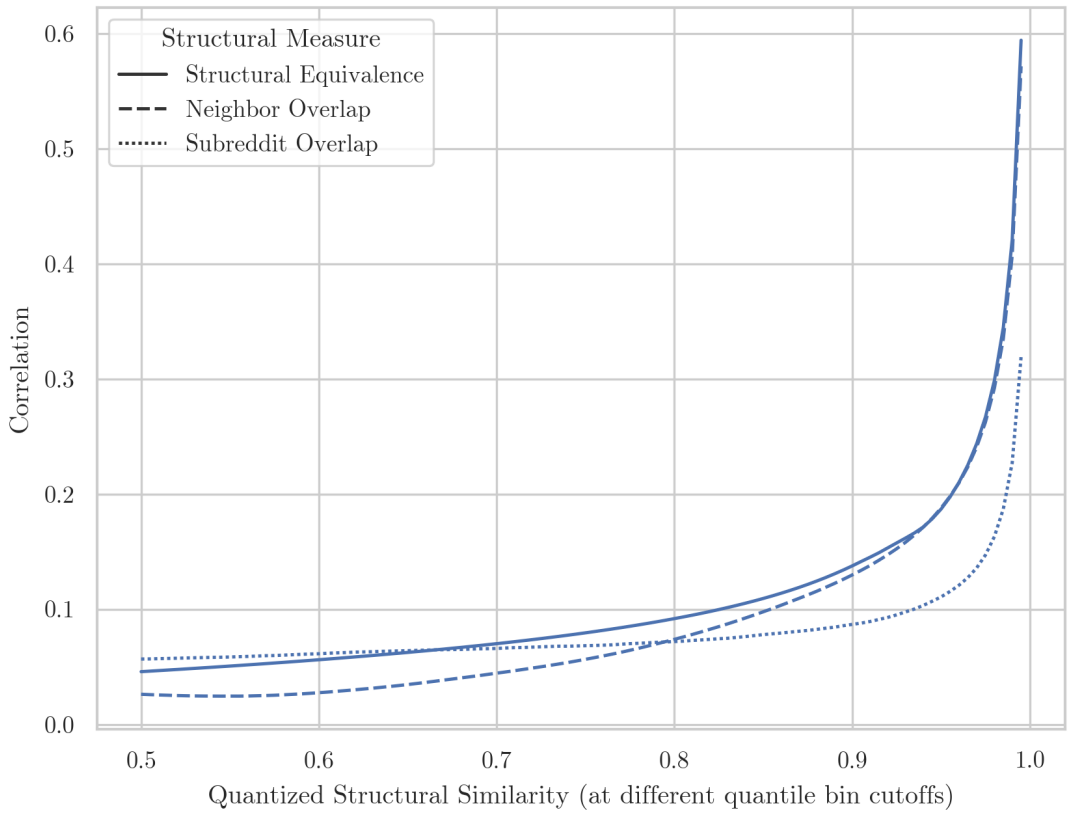
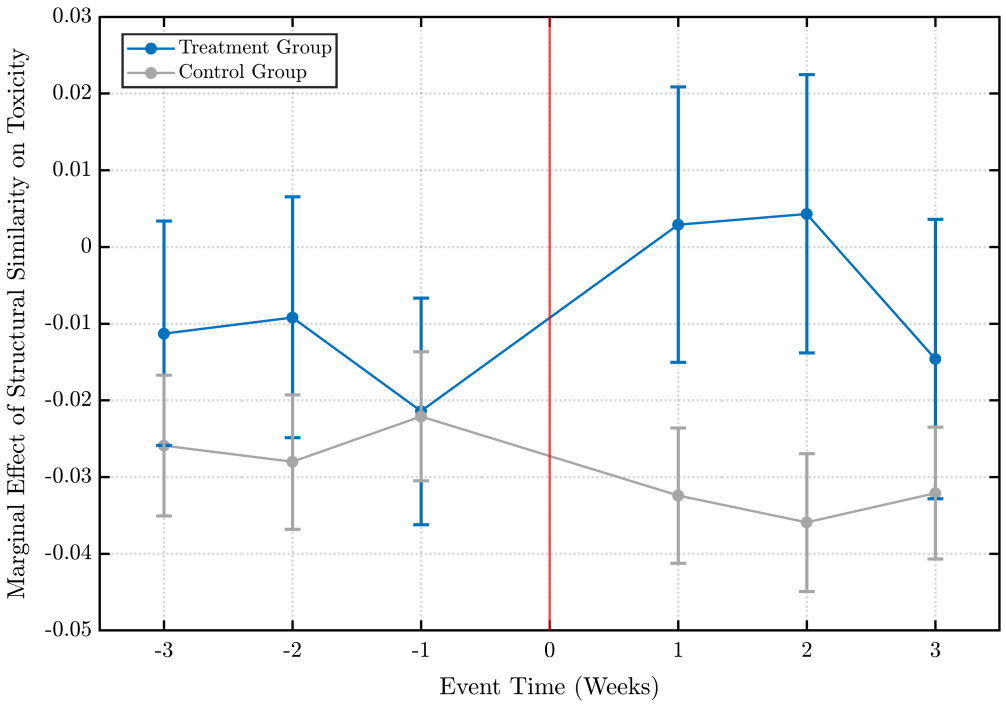


Figure 6: Event Study Representation of Difference-in-differences Analysis of a Shift in Group Polarization



Appendices

Additional Analyses and Robustness Checks

Table A1: Correlation Matrix: Structural Similarity measure at different values of hyperparameter q

q value	Mean	S.D.	**1**	**2**	**3**	**4**
(1) q = 0.5	0.596	0.220	–			
(2) q = 1	0.596	0.220	0.978***	–		
(3) q = 2	0.595	0.220	0.976***	0.977***	–	
(4) q = 3	0.594	0.220	0.974***	0.975***	0.976***	–

*Note: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$*

Table A2: OLS Regressions of Toxicity on Alternative Structural Similarity Measures and Semantic Similarity

Dependent Variable:	Toxicity							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Structural Similarity	-0.0077***							
	(0.0015)							
Eigenvector Similarity		0.0065						
		(0.0048)						
Indegree Similarity			0.0089					
			(0.0124)					
Outdegree Similarity				0.0088				
				(0.0080)				
Clustering Similarity					-0.0015			
					(0.0015)			
Subreddit Overlap						-0.0015		
						(0.0008)		
Thread Overlap							-0.0035*	
							(0.0017)	
Shared Ties								-0.0025***
								(0.0007)
Semantic Similarity	0.0317***	0.0317***	0.0317***	0.0317***	0.0317***	0.0317***	0.0317***	0.0317***
	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0011)
Fixed-effects								
Author	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Receiver	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subreddit	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	2,687,850	2,687,850	2,687,850	2,687,850	2,687,850	2,687,850	2,687,850	2,687,850
R^2	0.36762	0.36761	0.36761	0.36761	0.36761	0.36761	0.36761	0.36761

Two-way standard errors (clustered by sender and receiver) in parentheses.
*Note: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$*

Table A3: OLS Regression of Toxicity on Structural Similarity and Toxicity (With the Removal of Identified Bots)

Dependent Variable:	Toxicity					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Structural Similarity	-0.0829***	-0.0828***	-0.0182***	-0.0187***	-0.0146***	-0.0146***
	(0.0017)	(0.0017)	(0.0004)	(0.0004)	(0.0032)	(0.0032)
Semantic Similarity		-0.0060***		0.0329***		0.0114***
		(0.0013)		(0.0005)		(0.0025)
Fixed-effects						
Author			Yes	Yes	Yes	Yes
Receiver			Yes	Yes		
Week			Yes	Yes		
Subreddit			Yes	Yes	Yes	Yes
Dyad					Yes	Yes
Fit statistics						
Observations	24,820,417	24,820,417	24,820,417	24,820,417	24,820,417	24,820,417
R^2	0.00688	0.00691	0.23089	0.23164	0.90932	0.90932

Two-way standard errors (clustered by sender and receiver) in parentheses.
*Note: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$*

Table A4: OLS Regressions of the Sub-Components of Toxicity on Structural Similarity and Semantic Similarity

Dependent Variables:	Toxicity	SevereToxicity	IdentityAttack	Insult	Obscene	Threat
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Structural Similarity	-0.0194***	-0.0127***	-0.0121***	-0.0125***	-0.0137***	-0.0089***
	(0.0005)	(0.0005)	(0.0004)	(0.0004)	(0.0005)	(0.0004)
Semantic Similarity	0.0326***	0.0046***	0.0496***	0.0251***	0.0245***	0.0088***
	(0.0006)	(0.0005)	(0.0010)	(0.0007)	(0.0005)	(0.0004)
Fixed-effects						
Author	Yes	Yes	Yes	Yes	Yes	Yes
Receiver	Yes	Yes	Yes	Yes	Yes	Yes
Week	Yes	Yes	Yes	Yes	Yes	Yes
Subreddit	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	25,243,421	25,243,421	25,243,421	25,243,421	25,243,421	25,243,421
R^2	0.23480	0.19018	0.16847	0.21208	0.20710	0.14318

Two-way standard errors (clustered by sender and receiver) in parentheses.
*Note: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$*